Volumetric Correspondence Networks for Optical Flow

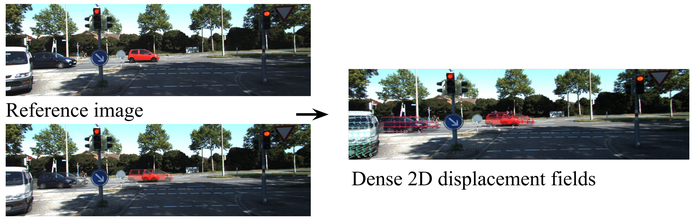
Gengshan Yang[1], Deva Ramanan[1,2]

Carnegie Mellon University[1], Argo AI[2]

Code available:
github.com/gengshan-y/VCN

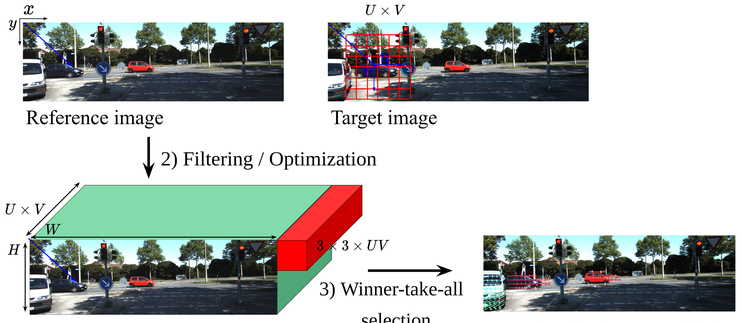
Introduction

Optical flow / visual correspondences



Related work: cost volume filtering

1) Correlation: $C(u,v,x,y) = \mathbf{F}_{ref}(x,y) \cdot \mathbf{F}_{tar}(x+u, y+v), \quad C(u,v,x,y) \in R^{U\times V\times H\times W}$



Contributions

·Efficient higher-dimension (4D) cost volume processing
- Separable volumetric filters: reduce computation and parameters
- Multi-channel cost volumes: capture multiple dimensions of pixel similarity
- Adaptive cost volumes allow networks to generalize across tasks: train and test a single network for both flow+stereo

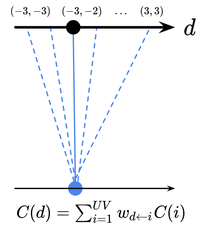
· Results
- SOTA accuracy on optical flow benchmarks
- Training converges in up to 10X fewer iterations than prior art

Approach

Key idea: 4D filters

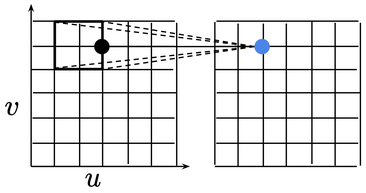
Prior method

1) reshape to $C(d,x,y) \in R^{UV\times H\times W}$
2) multi-channel 2D convs with $W \in R^{UV\times 3\times 3}$



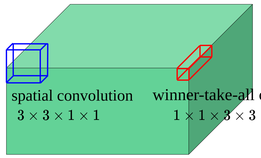
Ours

1) keep $C(u,v,x,y) \in R^{U\times V\times H\times W}$
2) 4D convs with $W \in R^{3\times 3\times 3\times 3}$



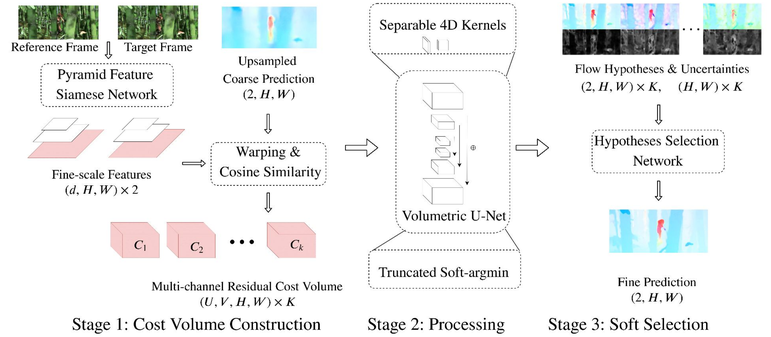
Separable filters

$K(\mathbf{u},\mathbf{x}) * C(\mathbf{u},\mathbf{x}) = K_{WTA}(\mathbf{u}) * \left[K_S(\mathbf{x}) * C(\mathbf{u},\mathbf{x})\right]$



Multi-channel 4D cost volumes

$$\begin{pmatrix} C_1(u,v,x,y) \\ C_2(u,v,x,y) \\ \vdots \\ C_K(u,v,x,y) \end{pmatrix} = \begin{pmatrix} \mathbf{F}_1(x,y)\cdot\mathbf{F'}_1(x+u,y+v) \\ \mathbf{F}_2(x,y)\cdot\mathbf{F'}_2(x+u,y+v) \\ \vdots \\ \mathbf{F}_K(x,y)\cdot\mathbf{F'}_K(x+u,y+v) \end{pmatrix}$$



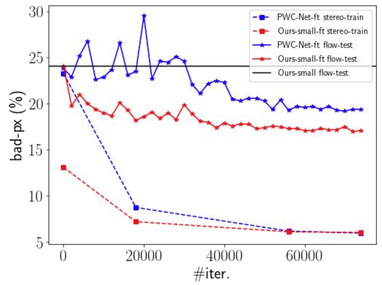
Experiments

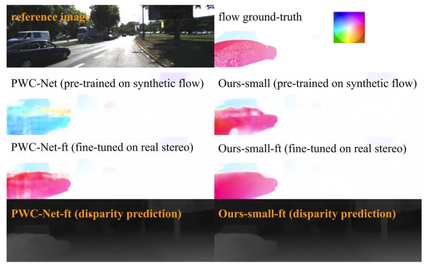
Benchmarks

Train dataset	Method	K-15-train Fl-epe↓	K-15-train Fl-all↓	K-15-test Fl-all↓	K-15-test D1-all↑ ↓	S-train (epe)↓ Clean	S-train (epe)↓ Final	S-test (epe)↓ Clean	S-test (epe)↓ Final	Param.	GFlops	Iter.
Pre-train: Chairs and Things	FlowNet2	10.08	30.0	-	-	**2.02**	**3.54**	3.96	6.02	162.5M	368.3	7100K
	PWC-Net	10.35	33.7	-	23.30	2.55	3.93	-	-	8.8M	101.6	1700K
	HD^3F	13.17	**24.0**	-	-	3.84	8.77	-	-	39.6M	174.8	-
	Ours-small	9.43	33.4	-	13.12	2.45	3.63	-	-	5.6M	41.0	220K
	Ours-full	**8.36**	25.1	-	**8.73**	2.21	3.62	-	-	6.2M	101.7	220K
Fine-tune: K(ITTI) or S(intel)	FlowNet2	(2.30)	(8.6)	11.48	-	(1.45)	(2.01)	4.16	5.74	162.5M	368.3	+500K
	PWC-Net+	(1.50)	(5.3)	7.72	9.17	(1.71)	(2.34)	3.45	4.60	8.8M	101.6	+750K
	LiteFlowNet2	(1.47)	(4.8)	7.74	-	(1.30)	(1.62)	3.45	4.90	-	-	-
	IRR-PWC	(1.63)	(5.3)	7.65_3	-	(1.92)	(2.51)	3.84	4.58	6.4M	-	+750K
	HD^3F	(1.31)	(4.1)	6.55_2	-	(1.87)	(1.17)	4.79	4.67	39.6M	174.8	-
	Ours-small	(1.41)	(5.5)	7.74	6.10	(1.84)	(2.44)	3.26	4.73	5.6M	41.0	+140K
	Ours-full	(1.16)	(4.1)	$\mathbf{6.30_1}$	**4.67**	(1.66)	(2.24)	$\mathbf{2.81_1}$	$\mathbf{4.40_1}$	6.2M	101.7	+140K

* Metrics: Fl-epe is the average end-point (L2) error of optical flow vectors; Fl-all and D1-all are percentage of flow/stereo predictions with error less than some threshold.

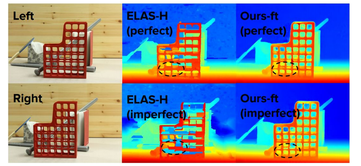
Generalize: 1D to 2D disparity





Application: stereo matching with imperfect rectification

Method	avgerr (px) perfect	avgerr (px) imperfect	inc.(%)
SGBM2	14.51	15.89	9.5
ELAS	9.89	11.79	19.2
PWC-Net	9.41	9.92	5.4
Ours	**9.03**	**8.79**	**-2.7**



Ablation

Method	EPE (px) ↓	GFlops	# Params.
DenseNet-2D	2.64	25.5	8.2
Ours-full-4D	2.30	52.5	1.83
Ours-sep-4D	2.31	23.4	1.78
Ours-final	**1.73**	28.5	2.94
Ours-$K=1$	2.05	27.8	2.94