RAFT:

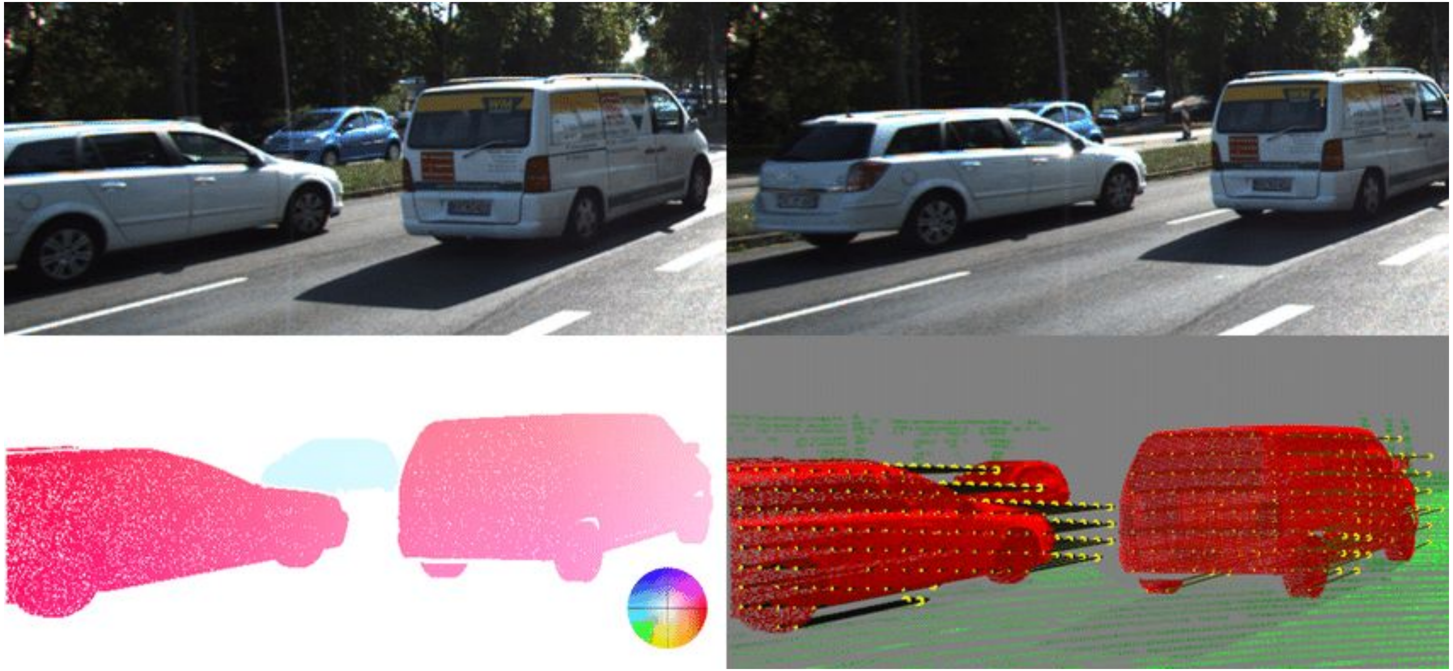
Recurrent **A**ll-Pairs **F**ield **T**ransforms for Optical Flow

Zachary Teed and Jia Deng
Princeton University

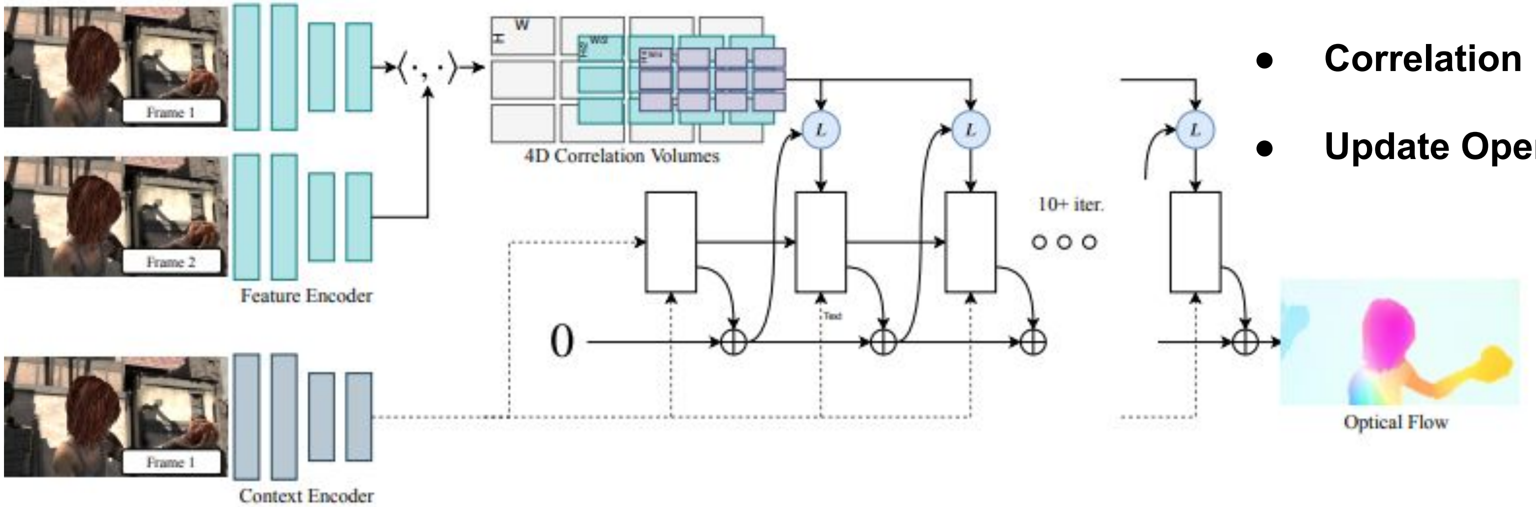
Presenter : Sungman Cho

Introduction

- Optical flow is the task of estimating per-pixel motion between video frames.



RAFT : architectures

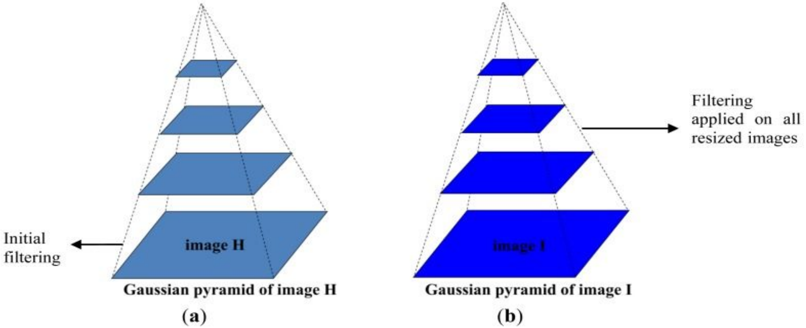


- **Feature Encoder**
- **Correlation Layer**
- **Update Operator**

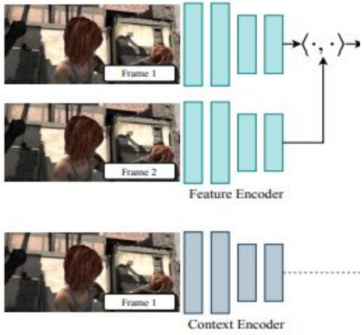
RAFT : novelty

- **RAFT** maintains and **updates a single fixed flow field at high resolution.** (This is different from the prevailing coarse-to-fine design)
    - The difficulty of recovering from errors at coarse resolutions, the tendency to **miss small fast-moving objects.**
    - The many training iterations typically required for training a multi-stage cascade.

[Coarse-to-fine]



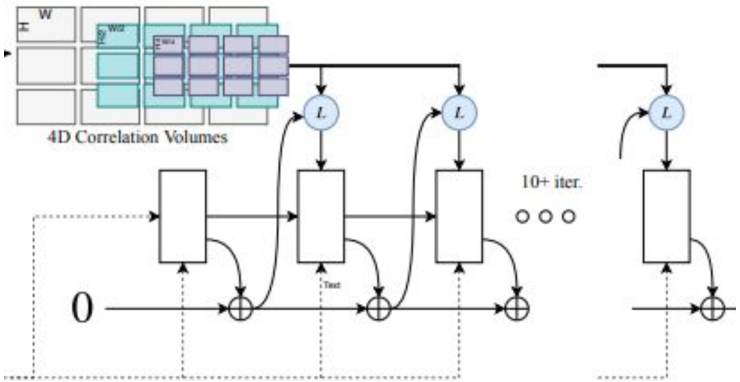
[RAFT]



RAFT : novelty

- **Update operator** of RAFT is **recurrent and lightweight.**
  (This is different from the prevailing coarse-to-fine design)

- The **update operator** has a **novel design**, which consists of a convolutional GRU that performs lookups on 4D multi-scale correlation volumes.



Contribution

- **State-of-the-art Accuracy**
  : On KITTI, F1-all error: 5.10

- **Strong generalization.**
  : trained only on synthetic data.

- **High efficiency**
  : 10fps at 1088x436 with 1080Ti

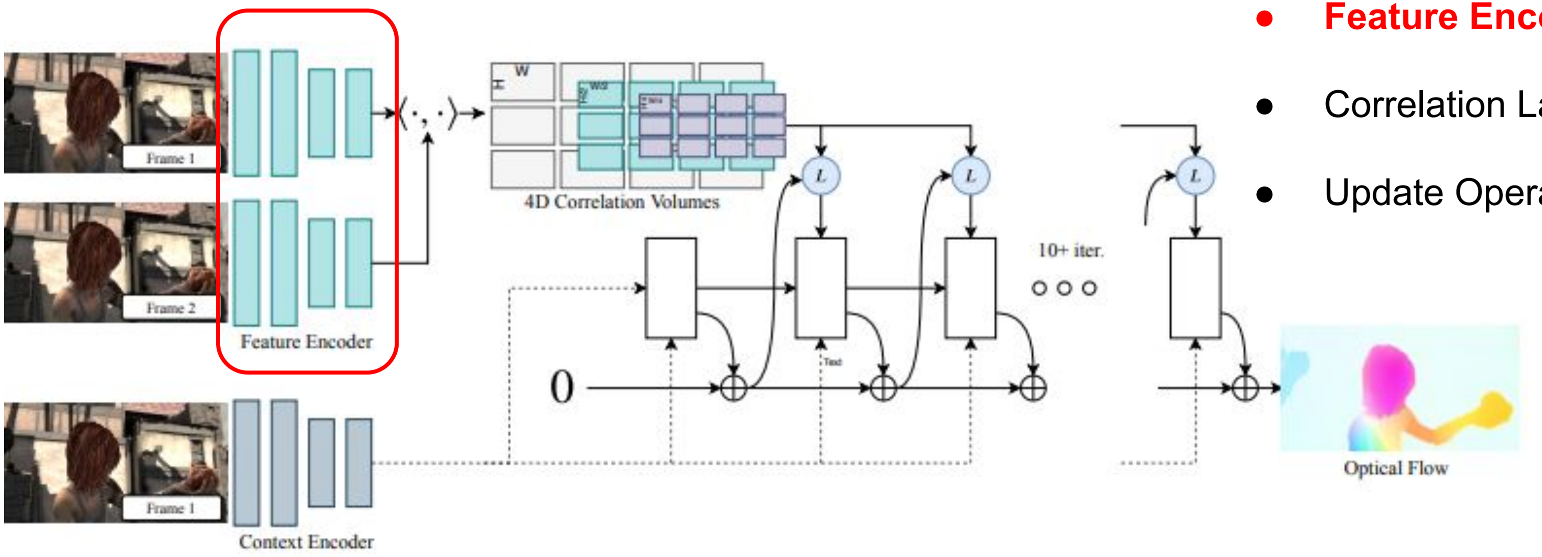
Methods

Approach

- Given a pair of consecutive RGB images, $I_1, I_2,$

  We estimate a dense displacement field $(\mathbf{f}^1, \mathbf{f}^2)$ which maps each pixel (u, v) in I_2 to its corresponding coordinates $(u', v') = (u + f^1(u), v + f^2(v))$

RAFT : Feature Extraction

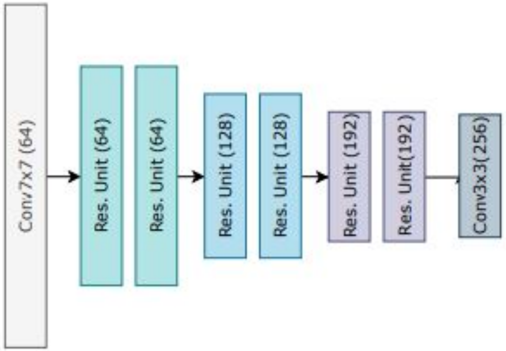


- **Feature Encoder**
- Correlation Layer
- Update Operator

RAFT : Feature Extraction

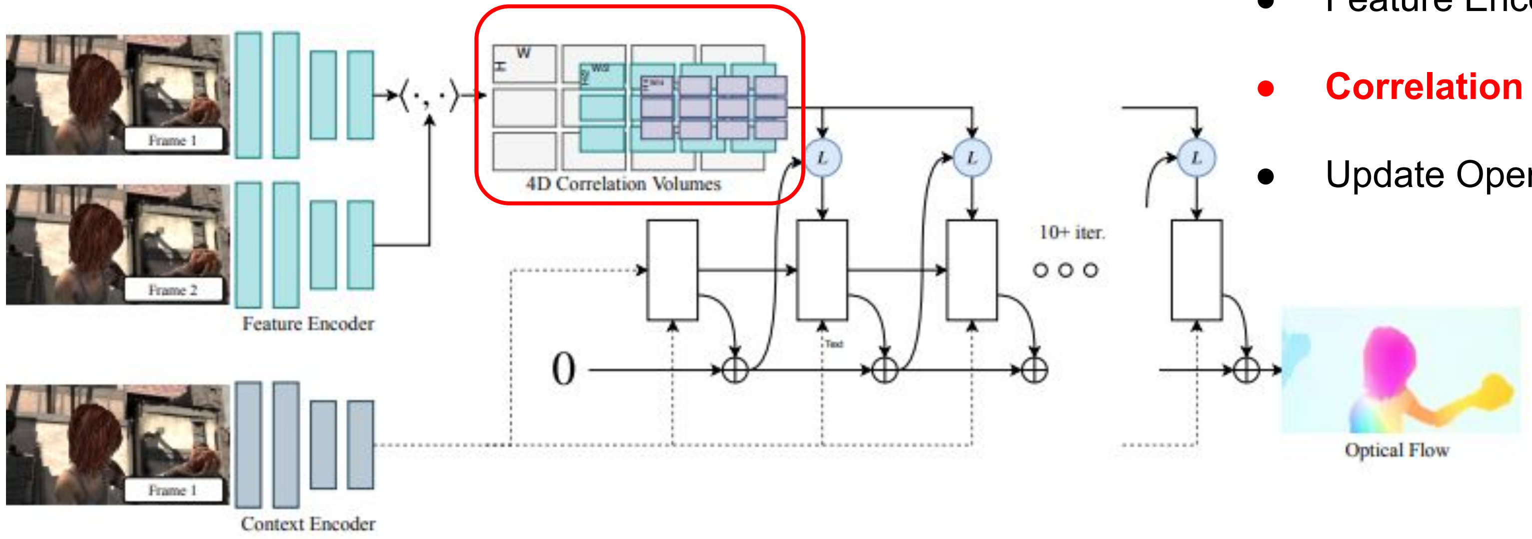
- The feature encoder consists of 6 residual blocks,
  2 at ½ resolution, 2 at ¼ resolution, 2 at ⅛ resolution.

$$g_\theta : \mathbb{R}^{H \times W \times 3} \mapsto \mathbb{R}^{H/8 \times W/8 \times D}$$



Feature / Context Encoder

RAFT : Computing Visual Similarity



- Feature Encoder
- **Correlation Layer**
- Update Operator

RAFT : Computing Visual Similarity

Given image features,

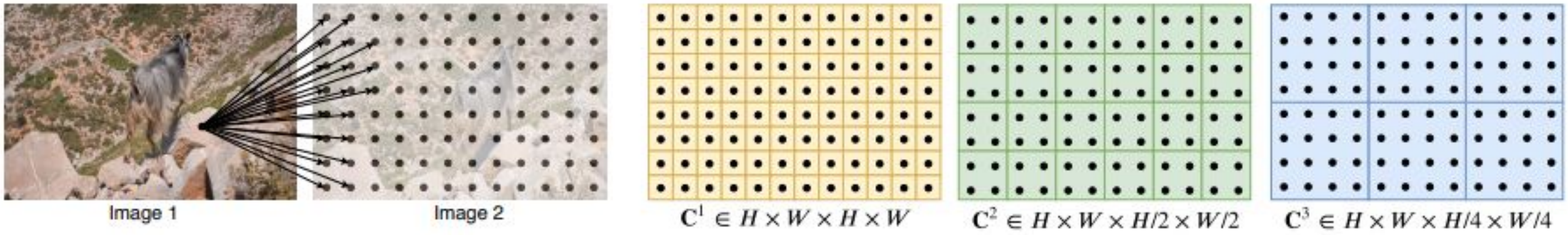
$$g_\theta(I_1) \in \mathbb{R}^{H \times W \times D} \quad g_\theta(I_2) \in \mathbb{R}^{H \times W \times D}$$

The correlation volume,

$$\mathbf{C}(g_\theta(I_1), g_\theta(I_2)) \in \mathbb{R}^{H \times W \times H \times W}, \qquad C_{ijkl} = \sum_h g_\theta(I_1)_{ijh} \cdot g_\theta(I_2)_{klh}$$

RAFT : Computing Visual Similarity

Correlation Pyramid

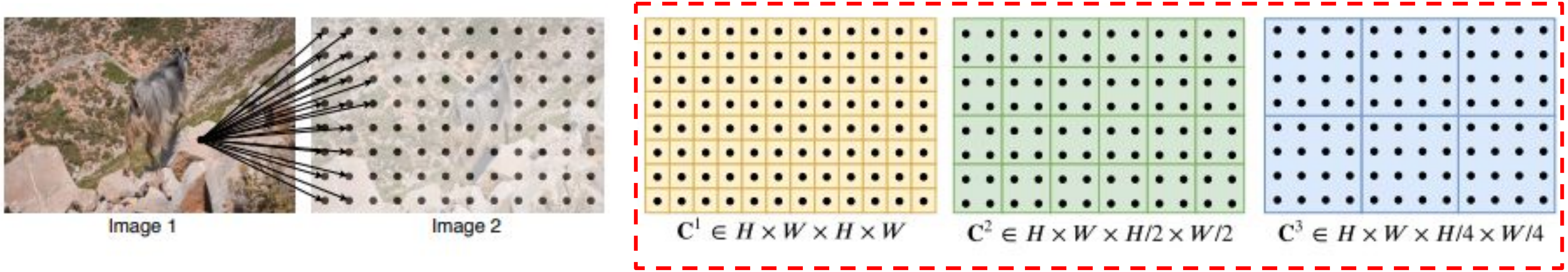


$$\mathbf{C}(g_\theta(I_1), g_\theta(I_2)) \in \mathbb{R}^{H \times W \times H \times W}, \qquad C_{ijkl} = \sum_h g_\theta(I_1)_{ijh} \cdot g_\theta(I_2)_{klh}$$

$\{\mathbf{C}^1, \mathbf{C}^2, \mathbf{C}^3, \mathbf{C}^4\}$ $\qquad$ $\mathbf{C}^k$ has dimensions $H \times W \times H/2^k \times W/2^k$.

RAFT : Computing Visual Similarity

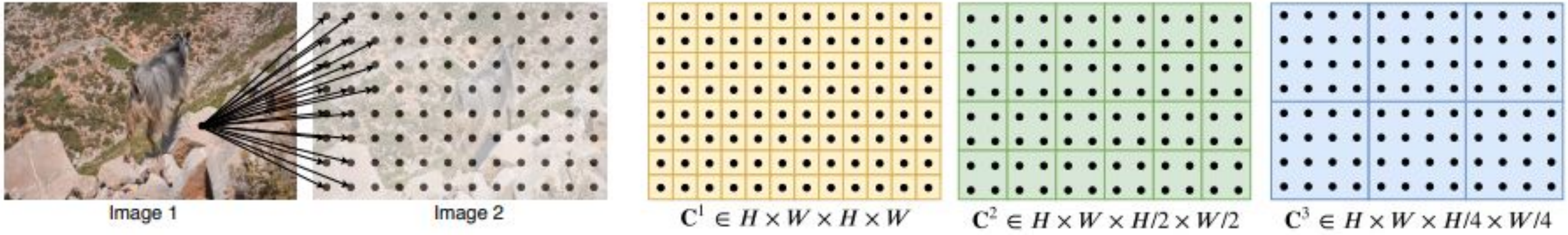
Correlation Lookup



$\mathcal{N}(\mathbf{x}')_r = \{\mathbf{x}' + \mathbf{dx} \mid \mathbf{dx} \in \mathbb{Z}^2, ||\mathbf{dx}||_1 \leq r\}$: indexing

RAFT : Computing Visual Similarity

Efficient Computation for High Resolution Images



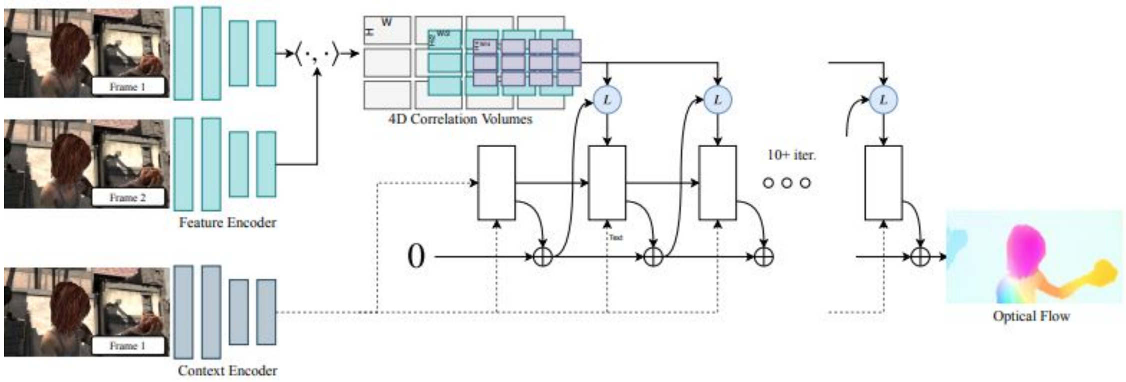
$$\mathbf{C}^m_{ijkl} = \underbrace{\boxed{\frac{1}{2^{2m}} \sum_p^{2^m} \sum_q^{2^m} \langle g^{(1)}_{i,j}, g^{(2)}_{2^m k+p, 2^m l+q} \rangle}}_{O(N^2)} = \underbrace{\boxed{\langle g^{(1)}_{i,j}, \frac{1}{2^{2m}} (\sum_p^{2^m} \sum_q^{2^m} g^{(2)}_{2^m k+p, 2^m l+q}) \rangle}}_{O(NM).}$$

RAFT : Computing Visual Similarity

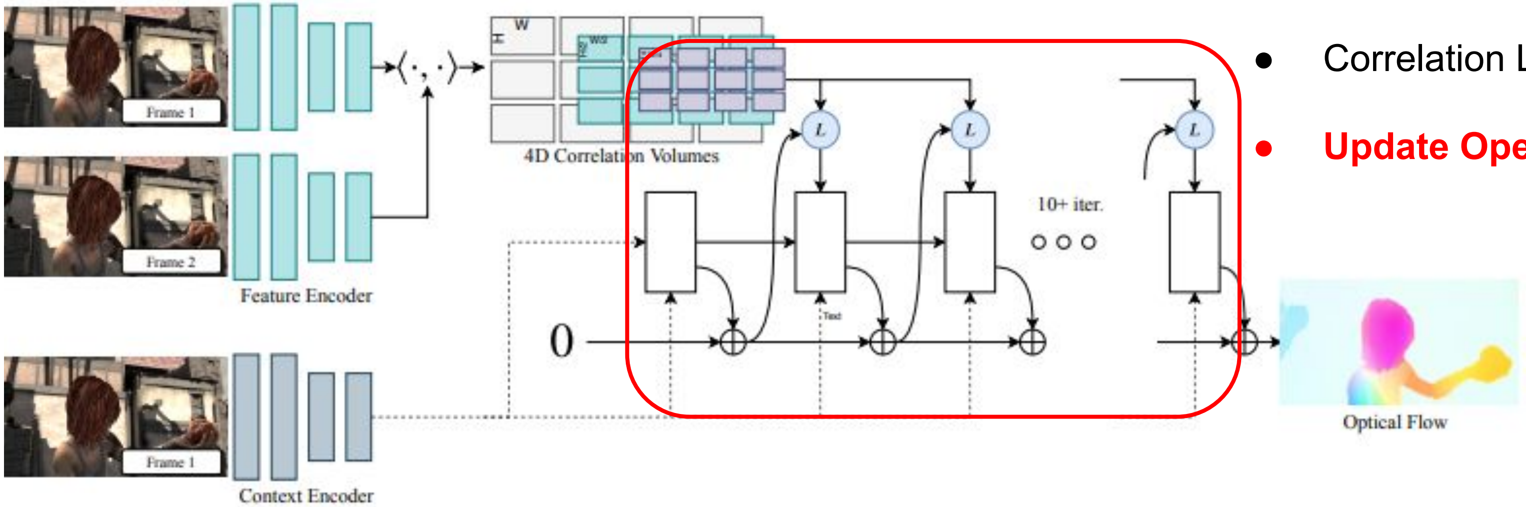
Efficient Computation for High Resolution Images

$$\mathbf{C}^m_{ijkl} = \boxed{\frac{1}{2^{2m}} \sum_p^{2^m} \sum_q^{2^m} \langle g^{(1)}_{i,j}, g^{(2)}_{2^m k+p, 2^m l+q} \rangle} = \boxed{\langle g^{(1)}_{i,j}, \frac{1}{2^{2m}} \left(\sum_p^{2^m} \sum_q^{2^m} g^{(2)}_{2^m k+p, 2^m l+q} \right) \rangle}$$

$$O(N^2) \qquad\qquad O(NM).$$

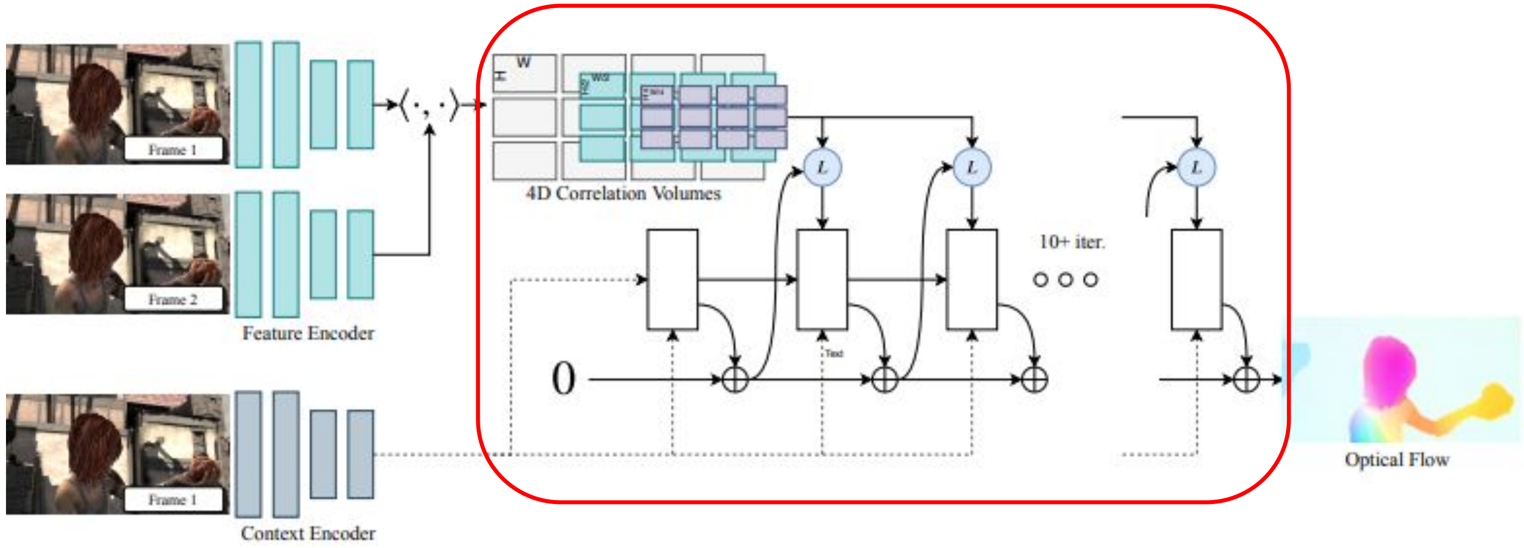


RAFT: Iterative Updates



- Feature Encoder
- Correlation Layer
- **Update Operator**

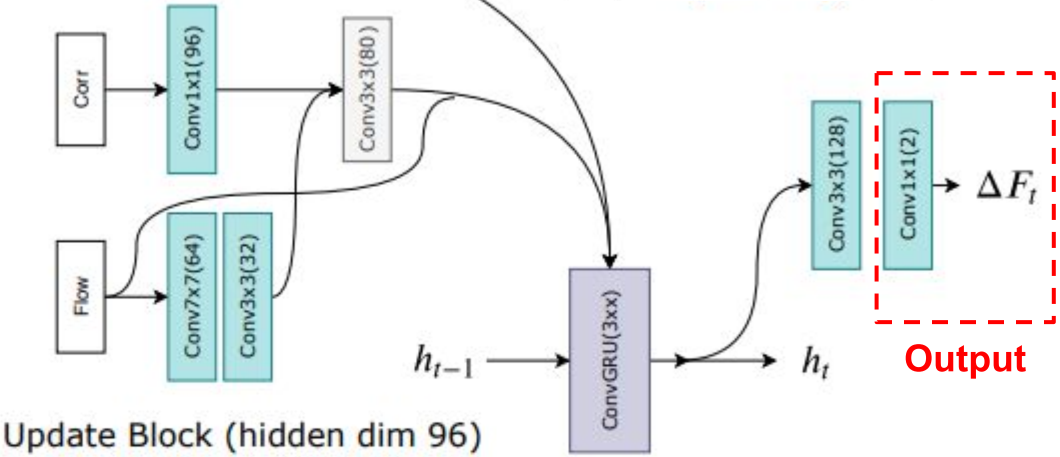
RAFT: Iterative Updates



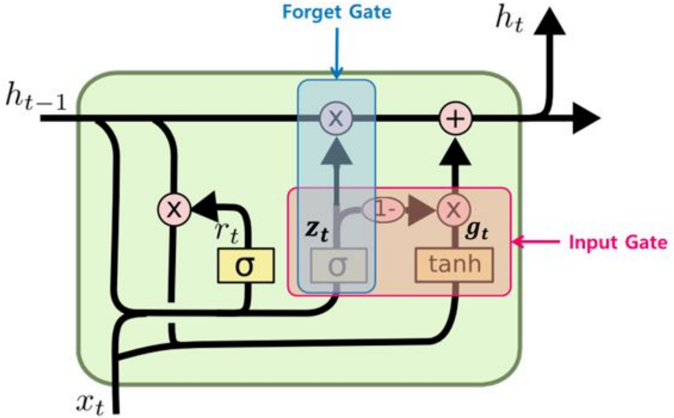
Inputs : flows, correlations, contexts, hidden-state

RAFT: Iterative Updates

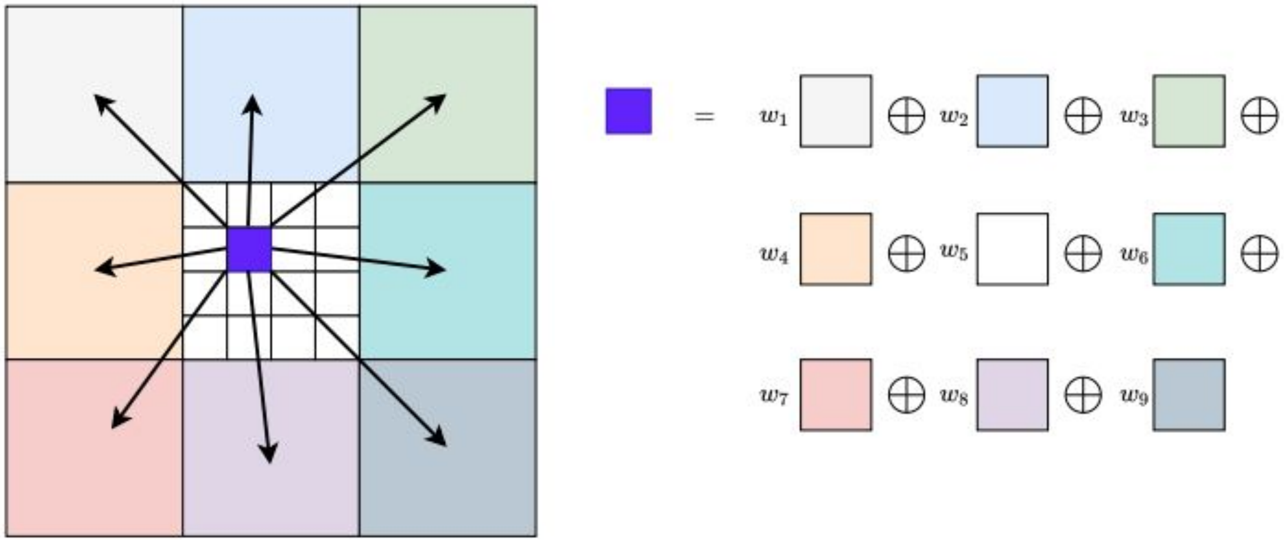
$$z_t = \sigma(\mathrm{Conv}_{3x3}([h_{t-1}, x_t], W_z))$$
$$r_t = \sigma(\mathrm{Conv}_{3x3}([h_{t-1}, x_t], W_r))$$
$$\tilde{h}_t = \tanh(\mathrm{Conv}_{3x3}([r_t \odot h_{t-1}, x_t], W_h))$$
$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$$



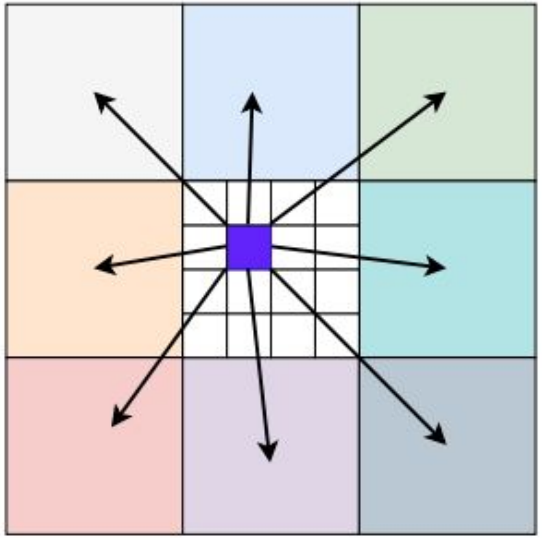
[Naive GRU]



RAFT: Convex Upsampling



RAFT: Convex Upsampling

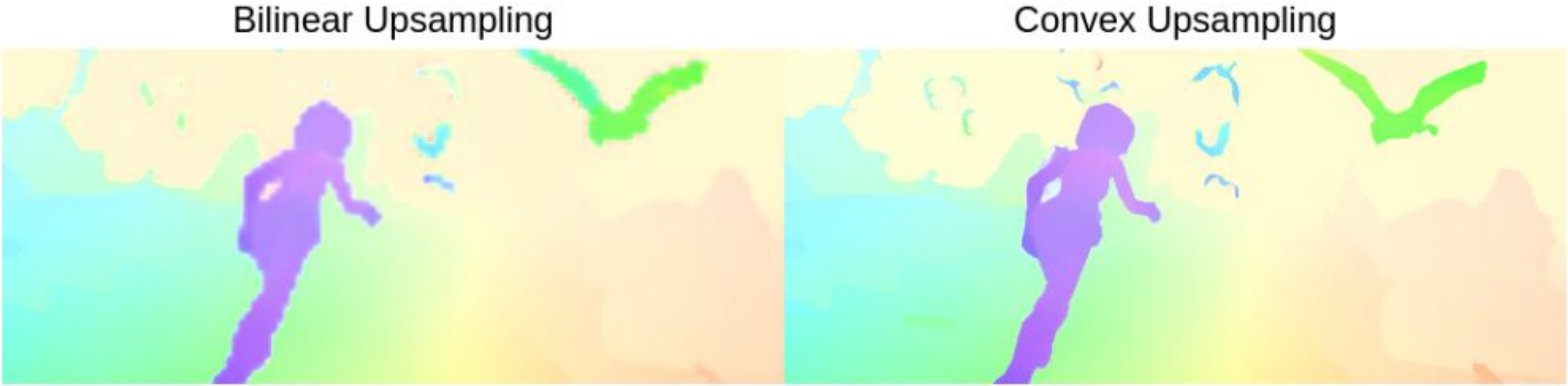


```python
def upsample_flow(self, flow, mask):
    """ Upsample flow field [H/8, W/8, 2] -> [H, W, 2] using convex combination """
    N, _, H, W = flow.shape
    mask = mask.view(N, 1, 9, 8, 8, H, W)
    mask = torch.softmax(mask, dim=2)

    up_flow = F.unfold(8 * flow, [3,3], padding=1)
    up_flow = up_flow.view(N, 2, 9, 1, 1, H, W)

    up_flow = torch.sum(mask * up_flow, dim=2)
    up_flow = up_flow.permute(0, 1, 4, 2, 5, 3)
    return up_flow.reshape(N, 2, 8*H, 8*W)
```

RAFT: Convex Upsampling



RAFT: Loss Function

$$\mathcal{L} = \sum_{i=1}^{N} \gamma^{N-i} ||\mathbf{f}_{gt} - \mathbf{f}_i||_1$$

i : sequence

Experiments

RAFT: Implementation Details

- Optimizer : AdamW , gradient clipping : [-1, 1]

- Flow updates : Sintel(32), KITTI(24).

- Training :

  FlyingThings (100k) → FlyingThings3D(100k) → FineTune (Sintel, KITTI-2015, HD1K)

RAFT: Results

Training Data	Method	Sintel (train)		KITTI-15 (train)		Sintel (test)		KITTI-15 (test)
		Clean	Final	F1-epe	F1-all	Clean	Final	F1-all
-	FlowFields[7]	-	-	-	-	3.75	5.81	15.31
-	FlowFields++[40]	-	-	-	-	2.94	5.49	14.82
S	DCFlow[47]	-	-	-	-	3.54	5.12	14.86
S	MRFlow[46]	-	-	-	-	2.53	5.38	12.19
C + T	HD3[50]	3.84	8.77	13.17	24.0	-	-	-
	LiteFlowNet[22]	2.48	4.04	10.39	28.5	-	-	-
	PWC-Net[42]	2.55	3.93	10.35	33.7	-	-	-
	LiteFlowNet2[23]	2.24	3.78	8.97	25.9	-	-	-
	VCN[49]	2.21	3.68	8.36	25.1	-	-	-
	MaskFlowNet[52]	2.25	3.61	-	23.1	-	-	-
	FlowNet2[25]	2.02	3.54[1]	10.08	30.0	3.96	6.02	-
	Ours (small)	2.21	3.35	7.51	26.9	-	-	-
	Ours (2-view)	**1.43**	**2.71**	**5.04**	**17.4**	-	-	-
C+T+S/K	FlowNet2 [25]	(1.45)	(2.01)	(2.30)	(6.8)	4.16	5.74	11.48
	HD3 [50]	(1.87)	(1.17)	(1.31)	(4.1)	4.79	4.67	6.55
	IRR-PWC [24]	(1.92)	(2.51)	(1.63)	(5.3)	3.84	4.58	7.65
	ScopeFlow[8]	-	-	-	-	3.59	4.10	6.82
	Ours (2-view)	(0.77)	(1.20)	(0.64)	(1.5)	**2.08**	**3.41**	**5.27**
C+T+S+K+H	LiteFlowNet2[2] [23]	(1.30)	(1.62)	(1.47)	(4.8)	3.48	4.69	7.74
	PWC-Net+[41]	(1.71)	(2.34)	(1.50)	(5.3)	3.45	4.60	7.72
	VCN [49]	(1.66)	(2.24)	(1.16)	(4.1)	2.81	4.40	6.30
	MaskFlowNet[52]	-	-	-	-	2.52	4.17	6.10
	Ours (2-view)	(0.76)	(1.22)	(0.63)	(1.5)	1.94	3.18	**5.10**
	Ours (warm-start)	(0.77)	(1.27)	-	-	**1.61**	**2.86**	-

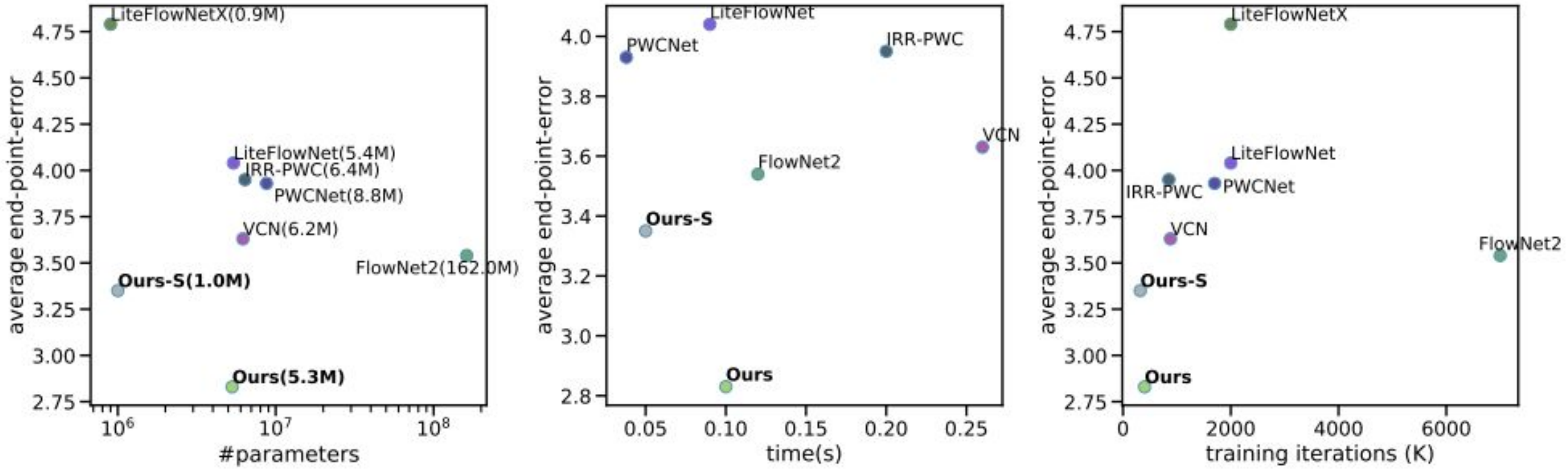
FlyingChairs(C) + FlyingThing(T) + S(Sintel) + K(KITTI) + H(HD1k)

RAFT: Ablation

Experiment	Method	Sintel (train) Clean	Sintel (train) Final	KITTI-15 (train) F1-epe	KITTI-15 (train) F1-all	Parameters
Reference Model (bilinear upsampling), Training: 100k(C) → 60k(T)						
Update Op.	ConvGRU	1.63	2.83	5.54	19.8	4.8M
	Conv	2.04	3.21	7.66	26.1	4.1M
Tying	Tied Weights	1.63	2.83	5.54	19.8	4.8M
	Untied Weights	1.96	3.20	7.64	24.1	32.5M
Context	Context	1.63	2.83	5.54	19.8	4.8M
	No Context	1.93	3.06	6.25	23.1	3.3M
Feature Scale	Single-Scale	1.63	2.83	5.54	19.8	4.8M
	Multi-Scale	2.08	3.12	6.91	23.2	6.6M
Lookup Radius	0	3.41	4.53	23.6	44.8	4.7M
	1	1.80	2.99	6.27	21.5	4.7M
	2	1.78	2.82	5.84	21.1	4.8M
	4	1.63	2.83	5.54	19.8	4.8M
Correlation Pooling	No	1.95	3.02	6.07	23.2	4.7M
	Yes	1.63	2.83	5.54	19.8	4.8M
Correlation Range	32px	2.91	4.48	10.4	28.8	4.8M
	64px	2.06	3.16	6.24	20.9	4.8M
	128px	1.64	2.81	6.00	19.9	4.8M
	All-Pairs	1.63	2.83	5.54	19.8	4.8M
Features for Refinement	Correlation	1.63	2.83	5.54	19.8	4.8M
	Warping	2.27	3.73	11.83	32.1	2.8M
Upsampling	Convex	1.43	2.71	5.04	17.4	5.3M
	Bilinear	1.60	2.79	5.17	19.2	4.8M
Inference Updates	1	4.04	5.45	15.30	44.5	5.3M
	3	2.14	3.52	8.98	29.9	5.3M
	8	1.61	2.88	5.99	19.6	5.3M
	32	1.43	2.71	5.00	17.4	5.3M
	100	1.41	2.72	4.95	17.4	5.3M
	200	1.40	2.73	4.94	17.4	5.3M

RAFT: Timing and Parameters

- Training: C+T / Test: Sintel.



RAFT: Results

Thank You.