RAFT: Recurrent All-Pairs Field Transforms for Optical Flow

CVPR 2020

AUTHOR：Zachary Teed, Jia Deng
REPORTER : 鄧仲恩

Outline

1. Introduction
2. Related Work
3. Approach
4. Experiments
5. Conclusion

Introduction

Introduction - Optical flow

1. Optical flow
2. Tradition (Full flow)

   a) not robust

   b) difficulties in hand-designing
3. Deep Learning(Flownet 2.0、Deepv2d、Liteflownet)

   a) directly predict flow

   b) faster

   c) better performance

Introduction - Recurrent All-Pairs Field Transforms

1. State-of-the-art accuracy

   a) F1-all error 6.1% -> 5.1%

   b) end-point-error 4.098 pixels -> 2.855 pixels

2. Strong generalization

   5.04 pixel end-point-error with only synthetic data.

3. High efficiency

   10 frames per second.

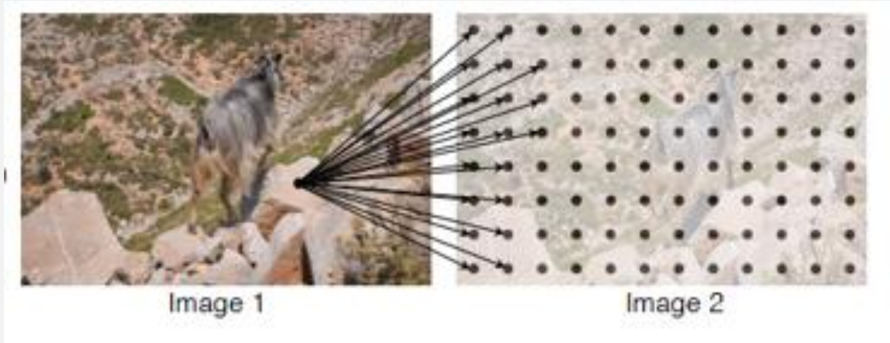
1	DEQ-Flow-H	13.0	3.76	Deep Equilibrium Optical Flow Estimation	2022
2	FlowFormer	14.7	4.09	FlowFormer: A Transformer Architecture for Optical Flow	2022
3	GMFlowNet	[illegible]	4.24	Global Matching with Overlapping Attention for Optical Flow Estimation	2022
4	SeparableFlow	15.9	4.60	Separable Flow: Learning Motion Cost Volumes for Optical Flow Estimation	2021
5	GMA	17.1	4.69	Learning to Estimate Hidden Motions with Global Motion Aggregation	2021
6	RAFT	17.4	5.04	RAFT: Recurrent All-Pairs Field Transforms for Optical Flow	2020
7	CRAFT	17.5	4.88	CRAFT: Cross-Attentional Flow Transformer for Robust Optical Flow	2022
8	SCV	19.3	6.80	Learning Optical Flow from a Few Matches	2021
9	MaskFlowNet	23.1		MaskFlownet: Asymmetric Feature Matching with Learnable Occlusion Mask	2020
10	HD3	24.0	13.17	Hierarchical Discrete Distribution Decomposition for Match Density Estimation	2018
11	VCN	25.1	8.36	Volumetric Correspondence Networks for Optical Flow	2019

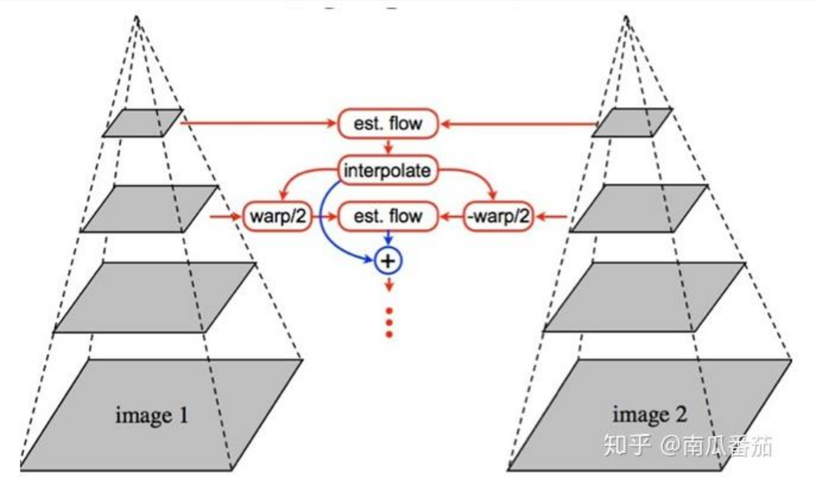
Related Work

Related Work - Optical Flow as Energy Minimization

$$E(u,v) = \iint [(I_x u + I_y v + I_t)^2 + \alpha^2 (\|\nabla u\|^2 + \|\nabla v\|^2)] dxdy$$

1. Horn and Schnuck use performing gradient steps.
2. TV-L1 use L1 data term and total variation regularization
3. Coarse-to-Fine
4. 4D cost volume





Related Work - Direct Flow Prediction

1. PWC-Net
2. LiteFlowNet
3. ScopeFlow
4. FlowNet

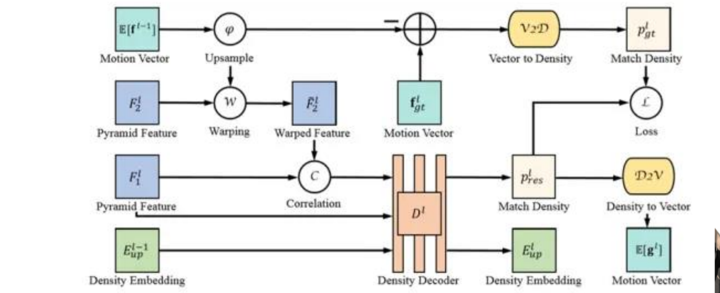
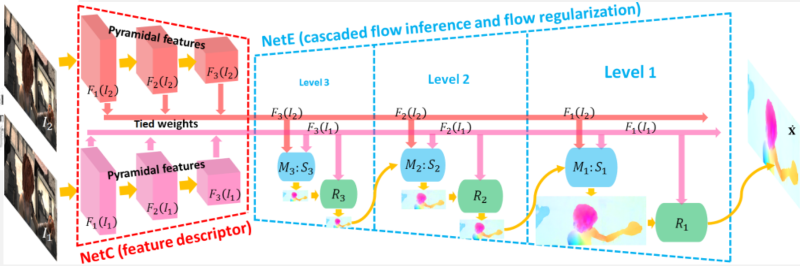


Figure 2: Overview of our architecture. The submodule at the l_{th} level is presented here. F^l and $\tilde{F}^l$ denotes the l_{th} level and warped pyramid features of image pair I. E^l_{up} denotes upsampled density embeddings between different levels as den connections. $\mathbf{f}^l$ and $\mathbf{g}^l$ denote motion vectors and p^l corresponds to match density. Their conversio[illegible] modules. For details please refer to our method part. This figure is best viewed in color.



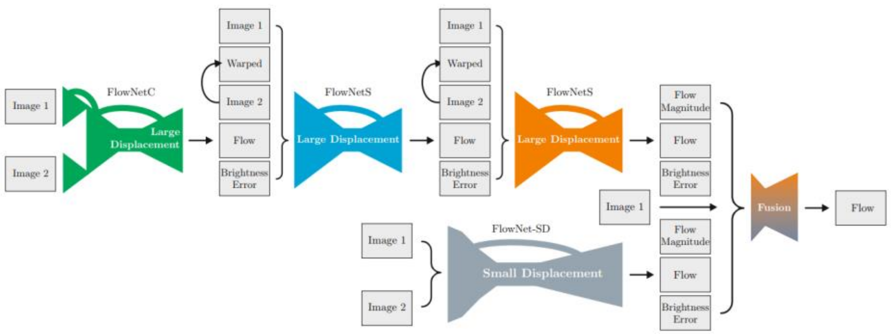
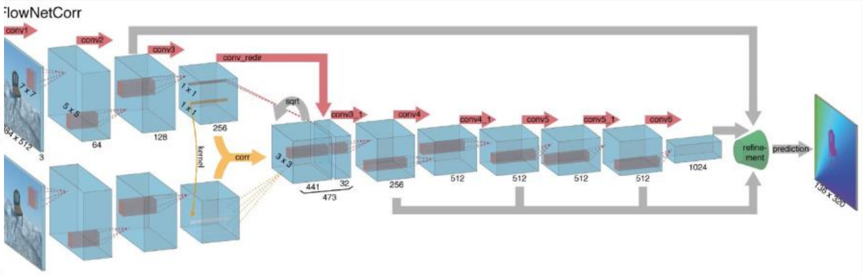


Figure 2. Schematic view of complete architecture: To compute large displacement optical flow we combine multiple FlowNets. Braces indicate concatenation of inputs. *Brightness Error* is the difference between the first image and the second image warped with the previously estimated flow. To optimally deal with small displacements, we introduce smaller strides in the beginning and convolutions between upconvolutions into the FlowNetS architecture. Finally we apply a small fusion network to provide the final estimate.



Related Work - Iterative Refinement for Optical Flow

1. FlowNet2.0
2. SpyNet , PWC-Net , LiteFlowNet, and VCN (coarse-to-fine pyramids)
3. IRR(shares weights)
4. TrellisNet and DEQ(LSTM)

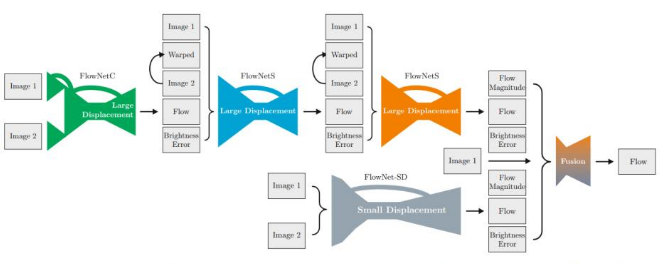
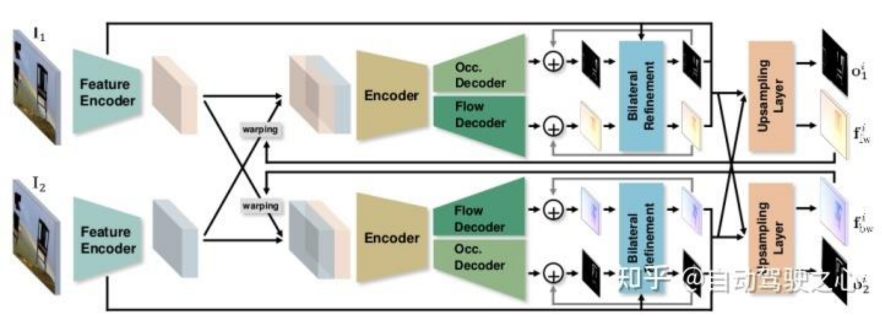
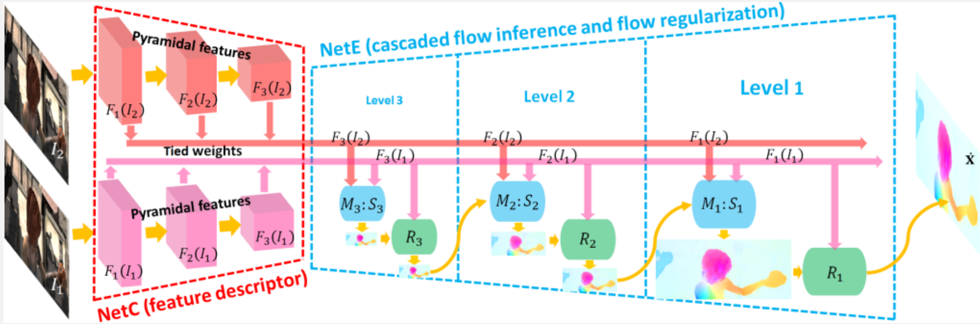


Figure 2. Schematic view of complete architecture: To compute large displacement optical flow we combine multiple FlowNets. Braces indicate concatenation of inputs. *Brightness Error* is the difference between the first image and the second image warped with the previously estimated flow. To optimally deal with small displacements, we introduce smaller strides in the beginning and convolutions between upconvolutions into the FlowNetS architecture. Finally we apply a small fusion network to provide the final estimate.

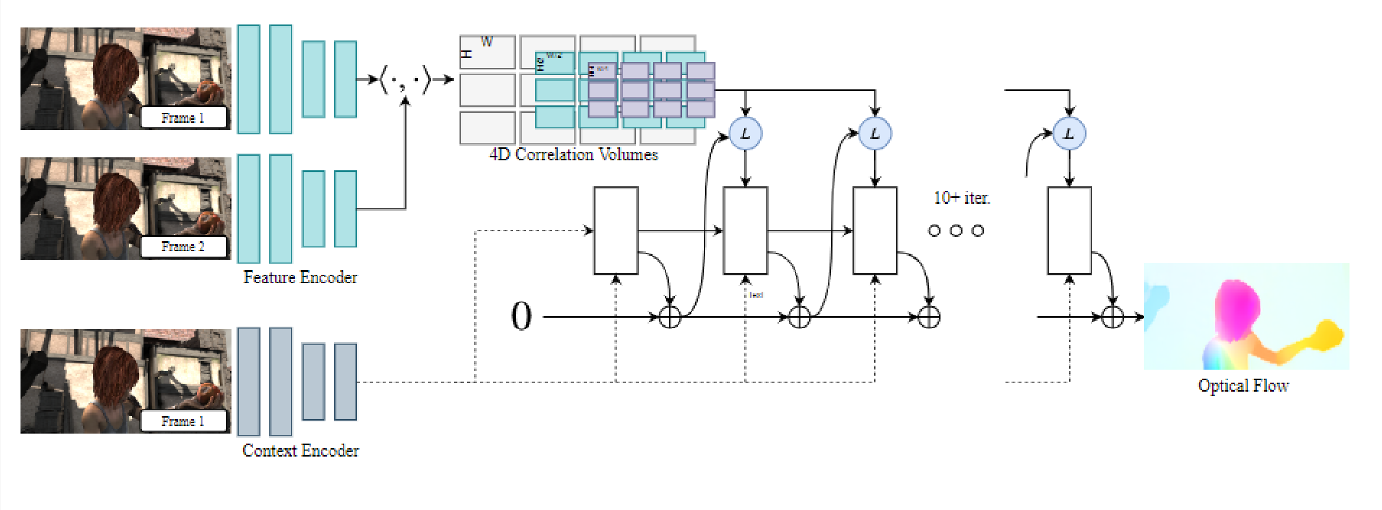




Approach

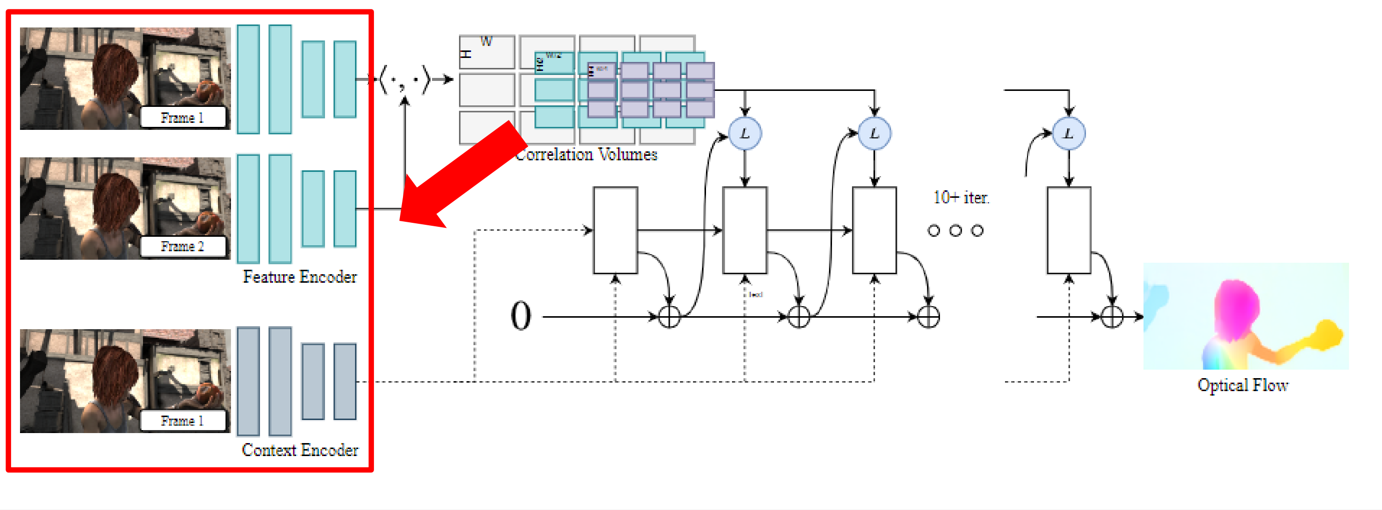
Approach

1. Feature Extraction
2. Computing Visual Similarity
3. Iterative Updates
4. Supervision



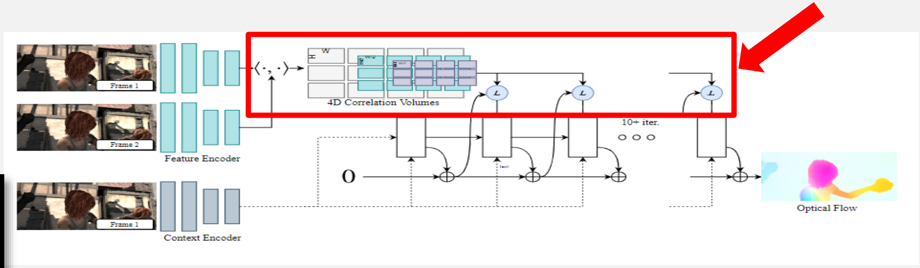
Approach - Feature Extraction

1. Image(H, W, 3) -> 1 / 8 features (H / 8, W / 8, 256)
2. 6 residual block



Approach - Computing Visual Similarity

1. .

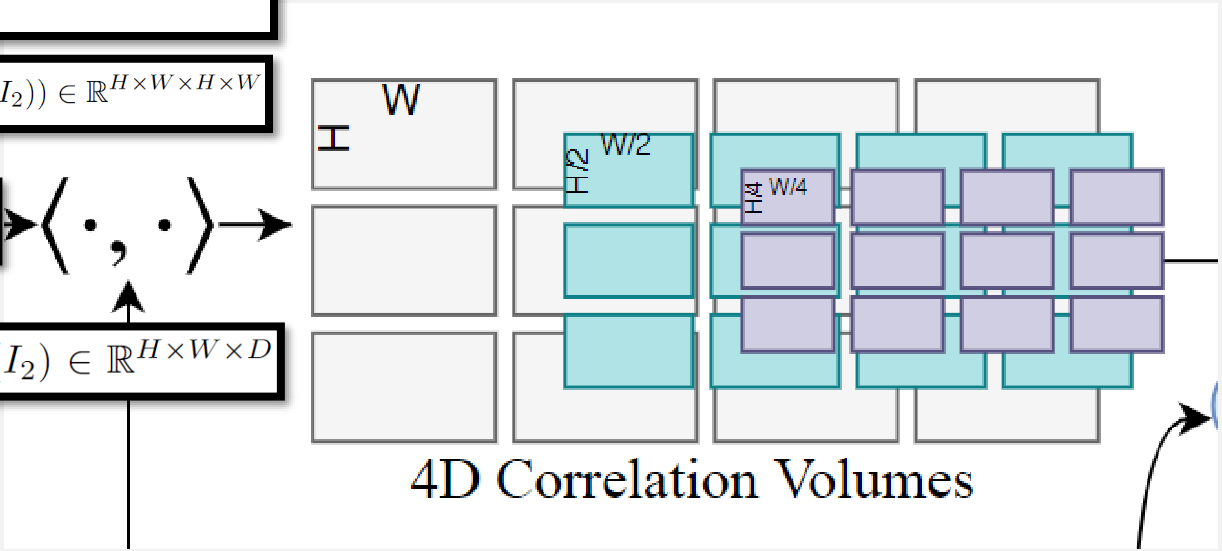


$$C_{ijkl} = \sum_h g_\theta(I_1)_{ijh} \cdot g_\theta(I_2)_{klh}$$

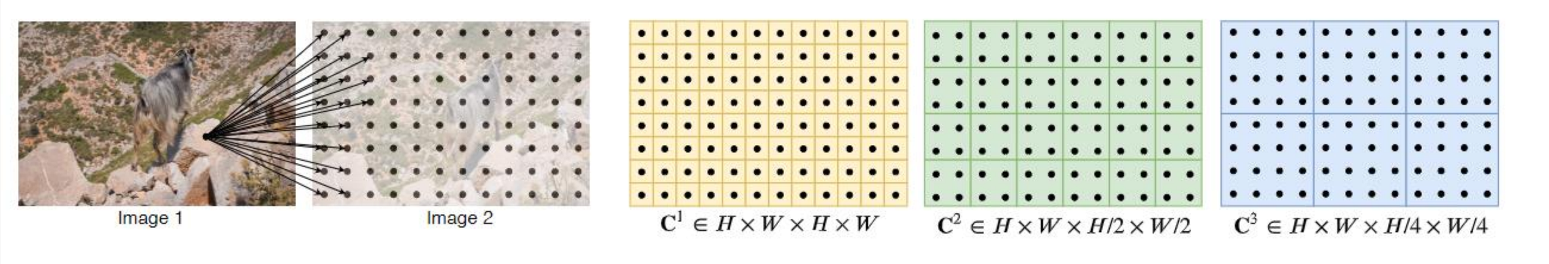
$$\mathbf{C}(g_\theta(I_1), g_\theta(I_2)) \in \mathbb{R}^{H \times W \times H \times W}$$

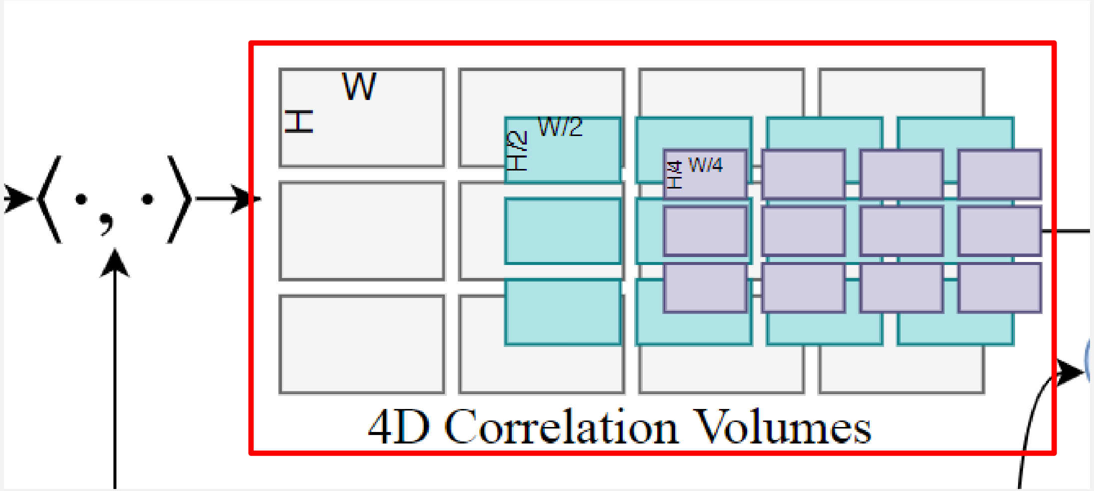
$$g_\theta(I_1) \in \mathbb{R}^{H \times W \times D}$$

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



Approach - Computing Visual Similarity(Pyramid)

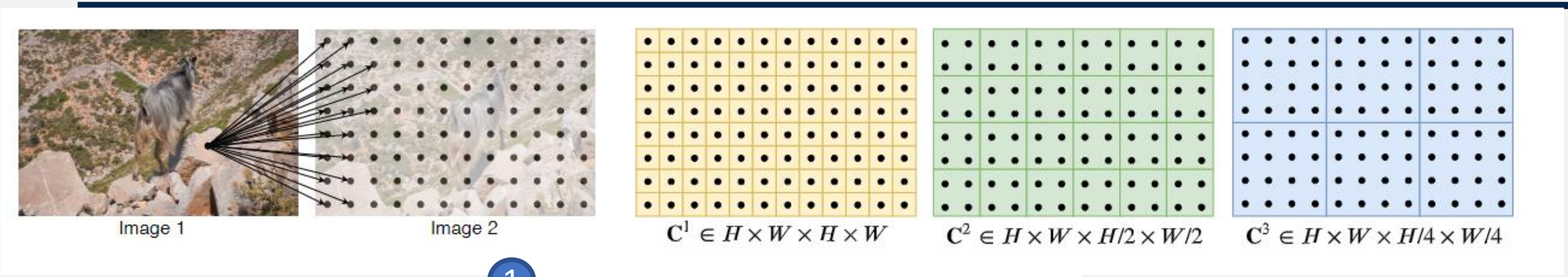




$$\{C^1, C^2, C^3, C^4\}$$

$$H \times W \times H/2^k \times W/2^k$$

Approach - Computing Visual Similarity(Lookup)



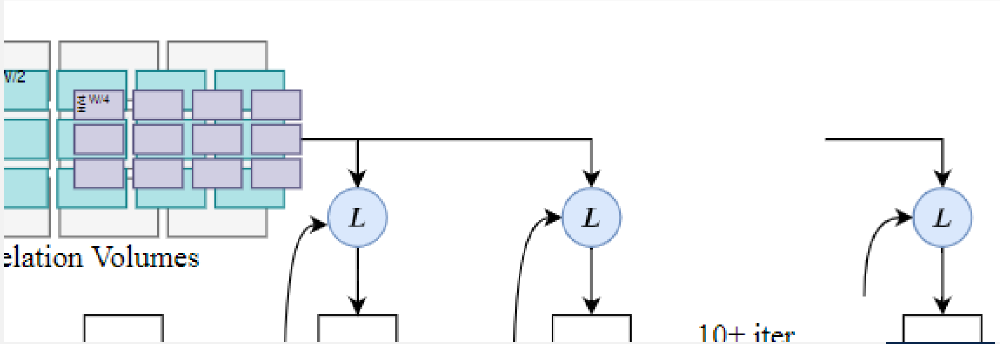
optical flow $(\mathbf{f}^1, \mathbf{f}^2)$

pixel $\mathbf{x} = (u, v)$

1 $\mathbf{x}' = (u + f^1(u), v + f^2(v))$

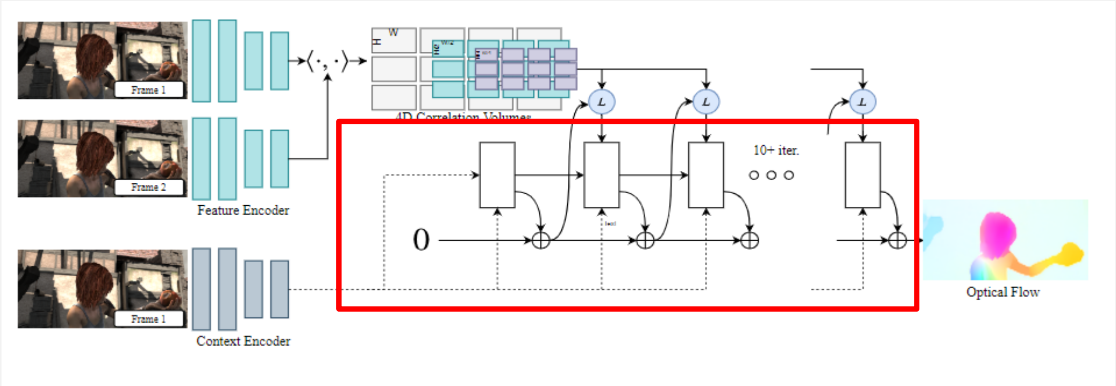
2 $\mathcal{N}(\mathbf{x}')_r = \{\mathbf{x}' + \mathbf{dx} \mid \mathbf{dx} \in \mathbb{Z}^2, ||\mathbf{dx}||_1 \leq r\}$

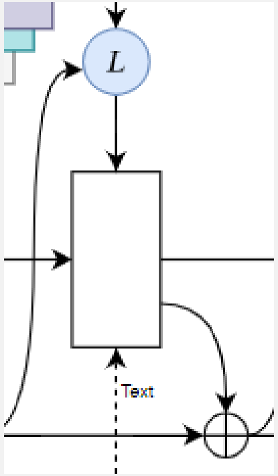
3 bilinear sampling->(batch_size, (R*2+1)^2*K, h, w)



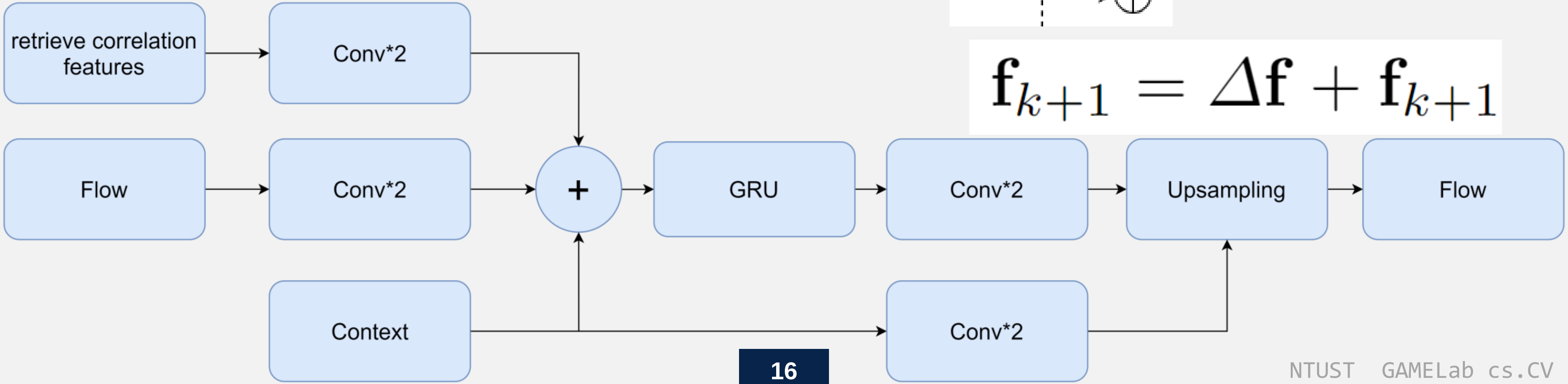

```
tensor([[[-3., -3.],
         [-3., -2.],
         [-3., -1.],
         [-3.,  0.],
         [-3.,  1.],
         [-3.,  2.],
         [-3.,  3.]],

        [[-2., -3.],
         [-2., -2.],
         [-2., -1.],
         [-2.,  0.],
         [-2.,  1.],
         [-2.,  2.],
         [-2.,  3.]],

        [[-1., -3.],
         [-1., -2.],
         [-1., -1.],
         [-1.,  0.],
         [-1.,  1.],
         [-1.,  2.],
         [-1.,  3.]],

        [[ 0., -3.],
         [ 0., -2.],
         [ 0., -1.],
         [ 0.,  0.],
         [ 0.,  1.],
         [ 0.,  2.],
         [ 0.,  3.]],

        [[ 1., -3.],
         [ 1., -2.],
         [ 1., -1.],
         [ 1.,  0.],
         [ 1.,  1.],
         [ 1.,  2.],
         [ 1.,  3.]],

        [[ 2., -3.],
         [ 2., -2.],
         [ 2., -1.],
         [ 2.,  0.],
         [ 2.,  1.],
         [ 2.,  2.],
         [ 2.,  3.]],

        [[ 3., -3.],
         [ 3., -2.],
         [ 3., -1.],
         [ 3.,  0.],
         [ 3.,  1.],
         [ 3.,  2.],
         [ 3.,  3.]]])
```

Approach - Iterative Updates

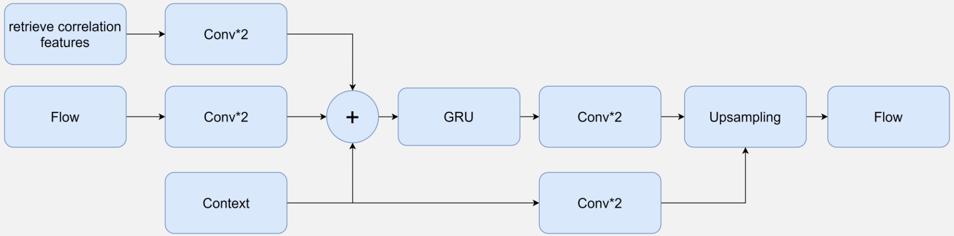


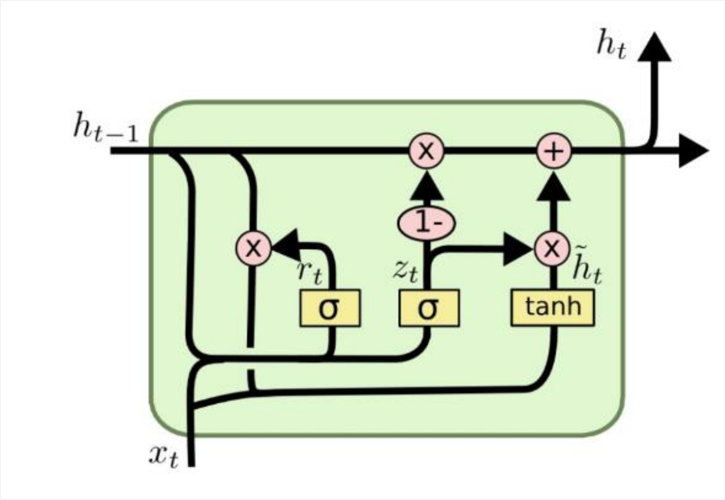


$$\mathbf{f}_{k+1} = \Delta \mathbf{f} + \mathbf{f}_{k+1}$$



Approach - Iterative Updates(GRU)





Conv1*5+Conv5*1

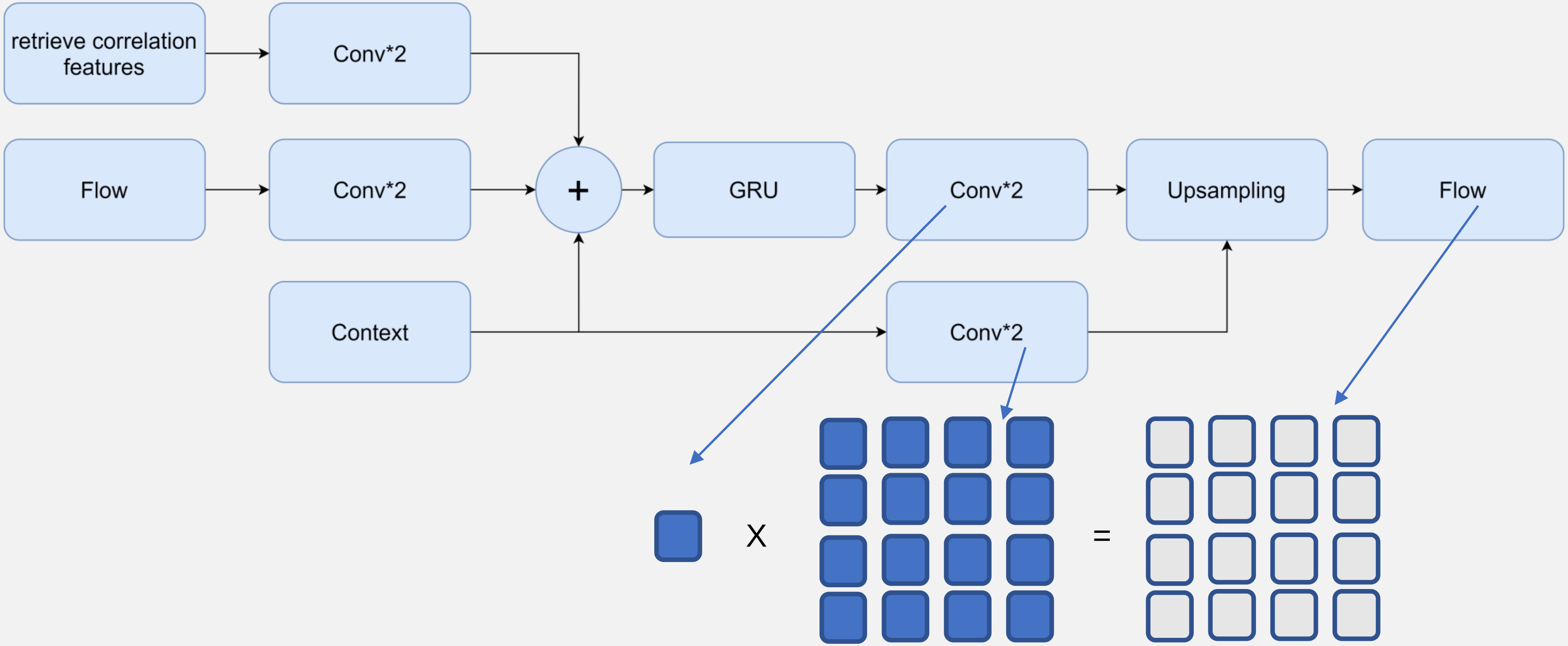
$$z_t = \sigma(\cancel{\mathrm{Conv}_{3x3}}([h_{t-1}, x_t], W_z))$$

$$r_t = \sigma(\cancel{\mathrm{Conv}_{3x3}}([h_{t-1}, x_t], W_r))$$

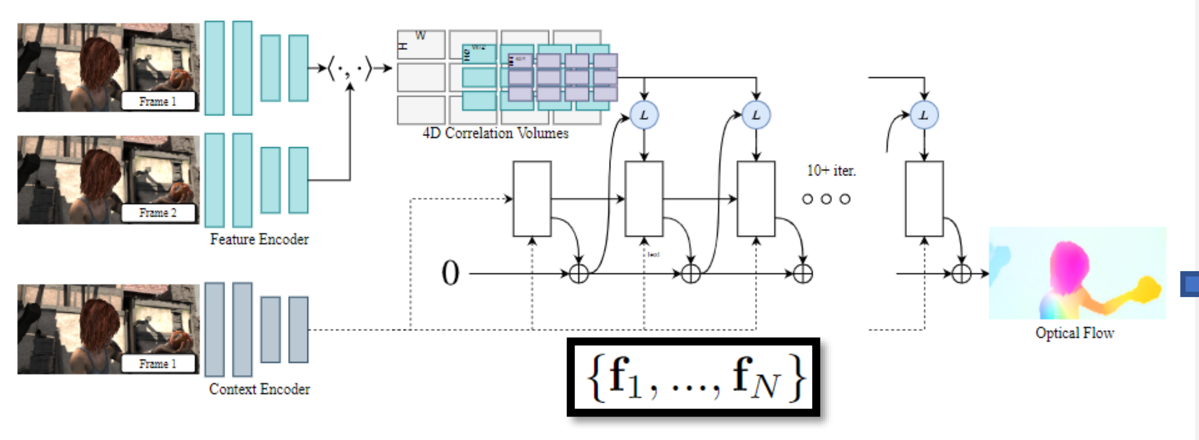
$$\tilde{h}_t = \tanh(\cancel{\mathrm{Conv}_{3x3}}([r_t \odot h_{t-1}, x_t], W_h))$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$$

Approach - Iterative Updates(Upsampling)



Approach - Supervision



$$\{\mathbf{f}_1, \ldots, \mathbf{f}_N\}$$

$$\gamma = 0.8$$

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

Experiments

Experiments

Dataset

1. Test on Sintel and KITTI.
2. Pretrain on FlyingChairs and FlyingThings.
3. Test on 1080p video from the DAVIS.

Details

1. 2080Ti GPUs.
2. AdamW
3. 32 flow updates on Sintel and 24 on KITTI
4. FlyingThings,100k iter, 12 batch size
5. FlyingThings3D,100k iter, 6 batch size
6. ……

Experiments - Sintel and KITTI

end-point-error

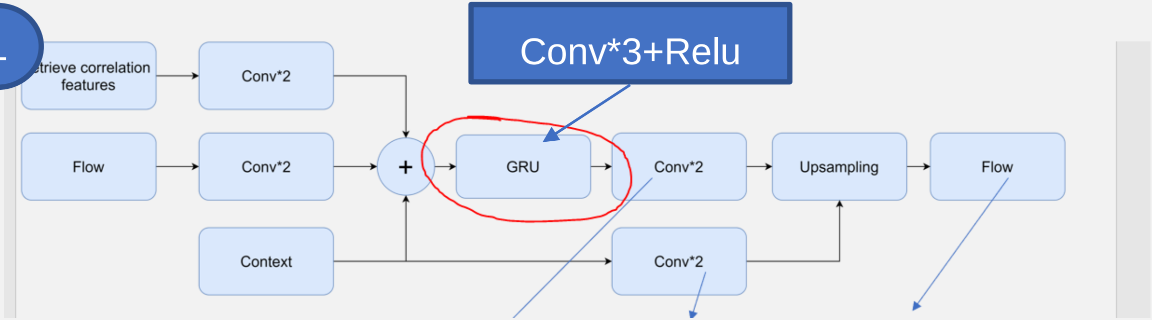
Training Data	Method	Sintel (train)		KITTI-15 (train)		Sintel (test)		KITTI-15 (test)
		Clean	Final	F1-epe	F1-all	Clean	Final	F1-all
-	FlowFields[7]	-	-	-	-	3.75	5.81	15.31
-	FlowFields++[40]	-	-	-	-	2.94	5.49	14.82
S	DCFlow[47]	-	-	-	-	3.54	5.12	14.86
S	MRFlow[46]	-	-	-	-	2.53	5.38	12.19
C + T	HD3[50]	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**	-

Experiments – Sintel and KITTI

Fig. 3: Flow predictions on the Sintel test set.

Experiments - Ablations

1

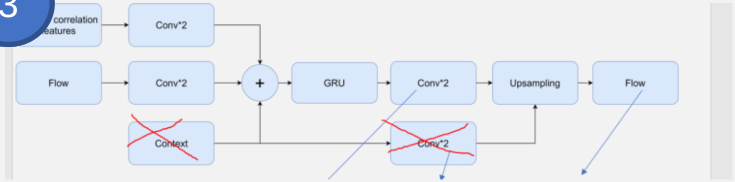


2

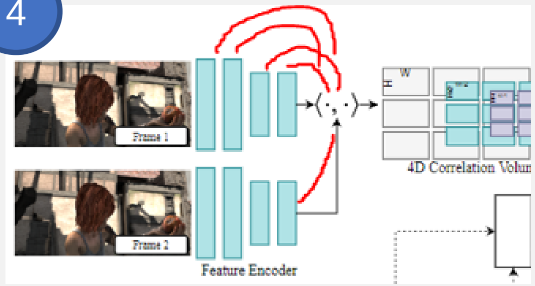
$= 0.8$

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

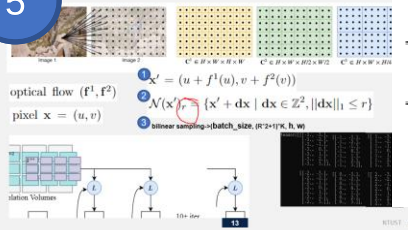
3



4



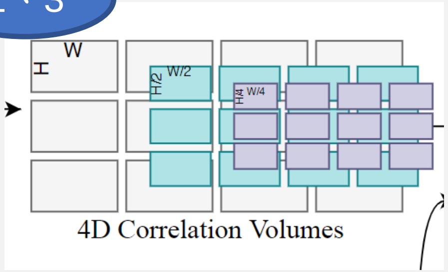
5



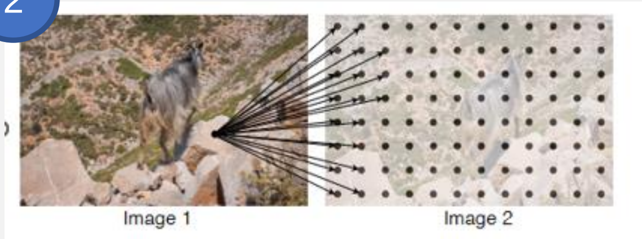
Experiment	Method	Sintel (train) Clean	Sintel (train) Final	KITTI-15 (train) F1-epe	KITTI-15 (train) F1-all	Parameters
Reference Model (bilinear upsampling), Training: 100k(C) → 60k(T)						
Update Op.	ConvGRU	1.63	2.83	5.54	19.8	4.8M
	Conv	2.04	3.21	7.66	26.1	4.1M
Tying	Tied Weights	1.63	2.83	5.54	19.8	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

Experiments - Ablations

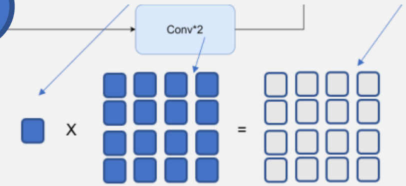
1、3



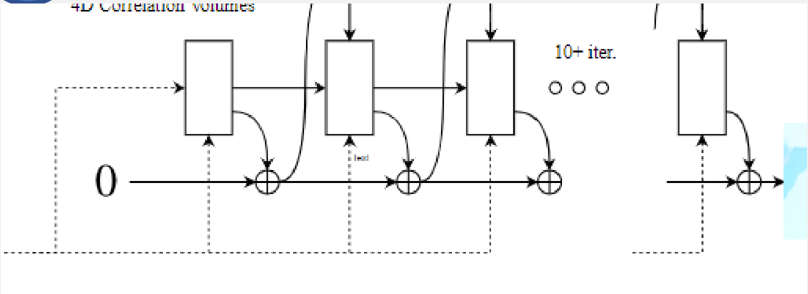
2



4

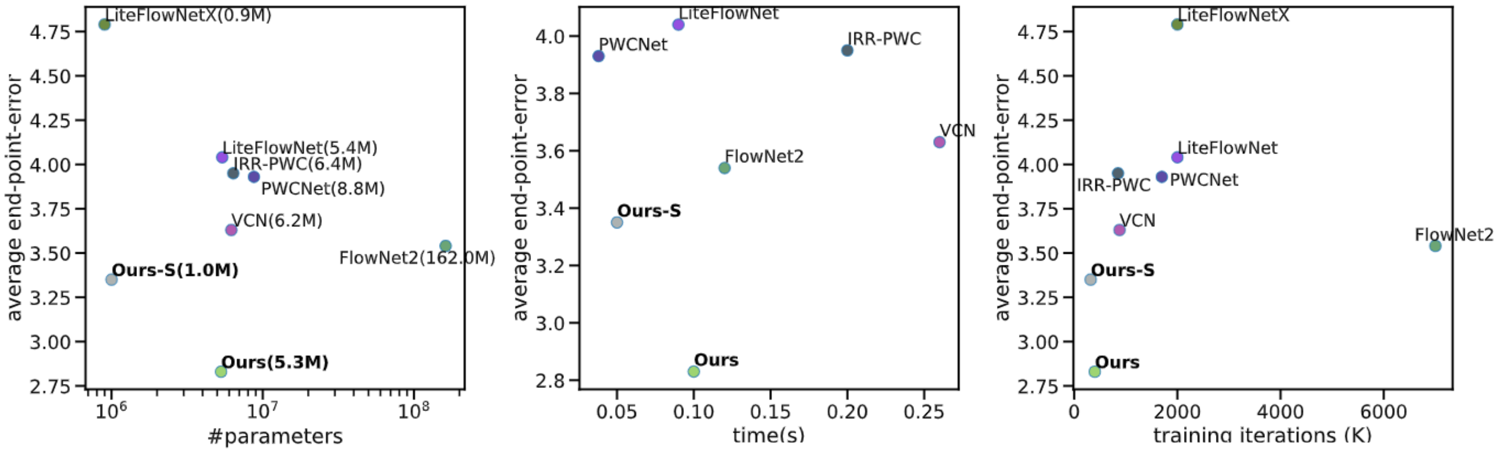


5



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
Reference Model (convex upsampling), Training: 100k(C) → 100k(T)						
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

Experiments - Timing and Parameter Counts



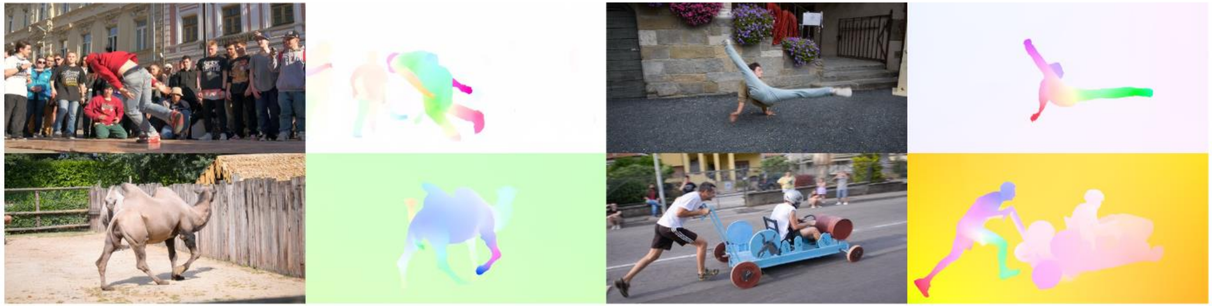


Fig. 6: Results on 1080p (1088x1920) video from DAVIS (550 ms per frame).

Conclusion

Conclusion

1. Proposed RAFT a new end-to-end trainable model for optical flow.
2. Achieves state-of-the-art accuracy across a diverse range of datasets.

1	DEQ-Flow-H	13.0	3.76	Deep Equilibrium Optical Flow Estimation	2022
2	FlowFormer	14.7	4.09	FlowFormer: A Transformer Architecture for Optical Flow	2022
3	GMFlowNet	15.4	4.24	Global Matching with Overlapping Attention for Optical Flow Estimation	2022
4	SeparableFlow	15.9	4.60	Separable Flow: Learning Motion Cost Volumes for Optical Flow Estimation	2021
5	GMA	17.1	4.69	Learning to Estimate Hidden Motions with Global Motion Aggregation	2021
6	RAFT	17.4	5.04	RAFT: Recurrent All-Pairs Field Transforms for Optical Flow	2020
7	CRAFT	17.5	4.88	CRAFT: Cross-Attentional Flow Transformer for Robust Optical Flow	2022
8	SCV	19.3	6.80	Learning Optical Flow from a Few Matches	2021
9	MaskFlowNet	23.1		MaskFlownet: Asymmetric Feature Matching with Learnable Occlusion Mask	2020
10	HD3	24.0	13.17	Hierarchical Discrete Distribution Decomposition for Match Density Estimation	2018
11	VCN	25.1	8.36	Volumetric Correspondence Networks for Optical Flow	2019

報告完畢 THE END

謝謝 Thank You

自評表

FAQ	完成
1. Paper的Input / Process / Output是什麼?	☑
2. Paper主要動機、目的、應用為何?如何應用到我們的研究?是否能應用到遊戲製作技巧上?	☑
3. Paper的貢獻?最重要的貢獻?價值?厲害在哪裡?強在哪裡?好在哪裡?優點?方法?想法?為什麼要這樣用?有什麼特別的?困難的地方在哪?為什麼難?跟Previous Work之間的差異是什麼?	☑
4. Paper的Framework?	☑
5. Paper中專有名詞的定義和意義和有哪些和正確發音?	☑
6. Paper中公式所代表的意義?物理意義?參數?為什麼要這樣定義的目的和原因?	☑
7. Paper的缺點?限制?哪邊可以改進?我們有沒有辦法發表一個方法解決?	x
8. Results的圖/表意義和如何閱讀?足夠嗎?數據為什能夠比其他的Paper好?是否有作弊?特別美化?為什麼實驗要這樣設計?結果的意義? 作者要讓我們知道什麼?如何驗證結果?如何從Results驗證Paper提到的優點、好處、貢獻?	☑
9. Paper是上什麼期刊?出處?Title的意思是什麼?被何人發表?	☑
10. 讀完Paper之後你有沒有辦法實現?用程式寫出來?	☑

What You've Learned

$$\mathbf{f}_{k+1} = \Delta \mathbf{f} + \mathbf{f}_{k+1}$$

Critical technique:

- Optical flow
- RAFT

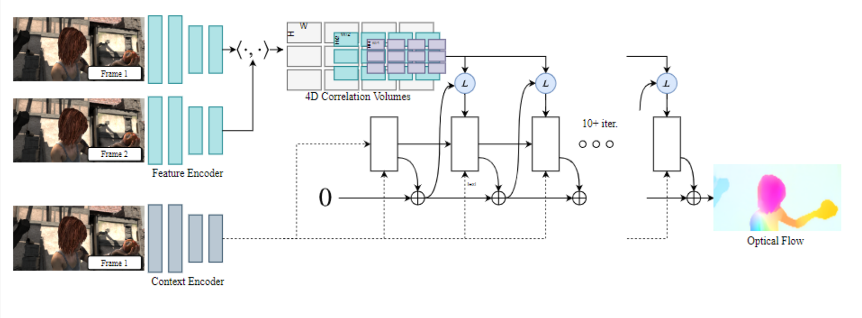
Good math expression:

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

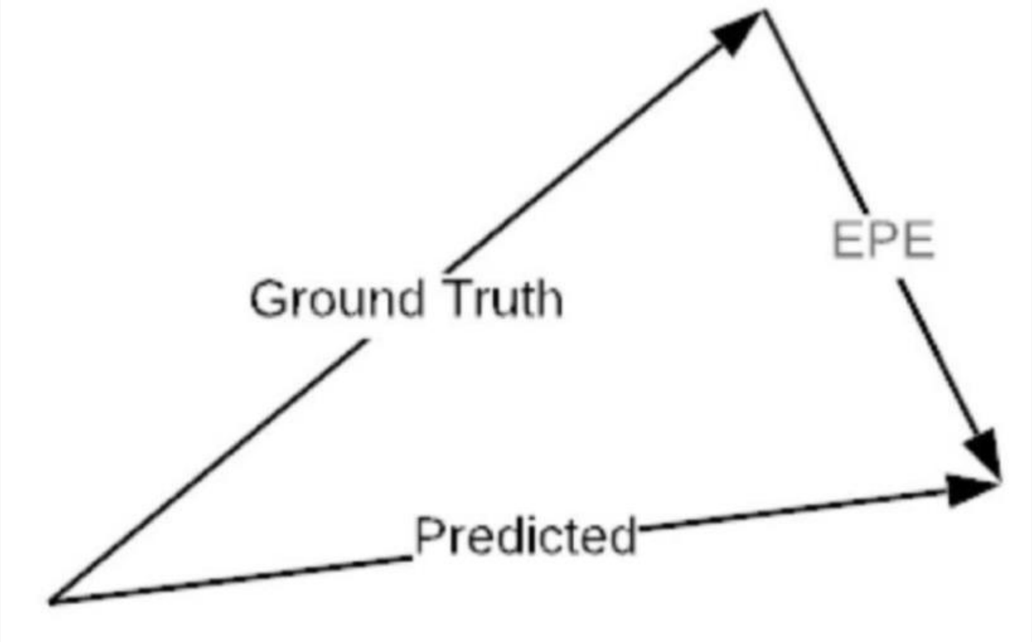
Recommended references:

All mentioned methods in this PPT.

Good experiment approach:



Other:


EPE
Ground Truth
Predicted