2025.01.08

Learning Optical Flow from Still Images

- Filippo Aleotti
- Matteo Poggi
- Stefano Mattoccia

컴퓨터전공 202474114

2025.01.08

Abstract & Introduction

- ❑ 저자
- ❑ 연구 배경
- ❑ 서론

0. Abstract

저자 및 개제년도

CVPR 2021

Department of Computer Science and Engineering (DISI)
University of Bologna, Italy

Filippo Aleotti

Matteo Poggi

Stefano Mattoccia

0. Abstract & Introduction

연구 배경

- **Optical Flow**
  - 비디오 프레임 간 픽셀 단위 움직임을 추정
  - 추적, 행동 인식 등 고차원 작업의 기반
  - Occlusions, motion blur, lack of texture 등의 문제가 있어 대량의 데이터셋이 필요

- 기존의 Optical flow 학습을 위한 데이터셋 확보
  - 합성 데이터셋Synthetic Dataset은 실제 데이터와 domain shift가 있어 일반화 성능이 떨어짐
    - Synthetic Dataset : 실제 데이터를 얻기 어려운 상황에서 모델을 학습시키기 위해 인위적으로 생성된 데이터
  - 라벨이 없는 원천 비디오 데이터는 라벨링 및 학습에 한계가 있음
  - 깊이 센서 LiDAR을 사용하더라도 수동 조정이 필수적이기 때문에 비효율적

0. Abstract & Introduction

제안

- 본 논문에서는 Depthstillation이라는 새로운 프레임워크를 제시
  - 단일 실제 이미지에서 깊이 추정 네트워크를 사용, 해당 장면의 3D 포인트 클라우드 사용
  - 가상 카메라를 이동시켜 새로운 뷰를 생성, 이에 대응하는 Optical Flow 필드 생성
  - 라벨이 없는 비디오나 합성 데이터셋에 의존하지 않고 대량의 Ground Truth데이터를 확보 가능

0. Abstract & Introduction

제안

2025.01.08

II Related Work & Method

- Optical Flow
- Self supervised Optical Flow
- Single image depth estimation
- View Synthesis

2. Related Work

Optical flow – Energy Minimization Models

에너지 최소화 모델 Energy Minimization Models

- 초기 Optical Flow 연구는 Variational Framework를 사용하여 에너지 최적화를 통해 픽셀의 움직임을 계산
  - **데이터 항 Data Term**: 두 프레임 간 픽셀의 RGB값 일치율
  - **정규화 항 Regularization Term**: 움직이는 픽셀의 주변 픽셀 또한 비슷한 방향으로 움직인하는 가정 >> 움직임의 부드러

- **장점**
  - 추정된 픽셀 움직임의 부드러움 향상

- **한계점**
  - 이러한 최적화 기반 방법은 계산 속도가 느리고 대규모 데이터에서 비효율적

2. Related Work

Optical flow – Deep Learning based

Deep Learning-based Optical Flow

- **FlowNet** (Dosovitskiy, 2015) >> Optical Flow를 예측하기 위한 최초의 **CNN 모델**
  - 합성 데이터셋(FlyingChairs)을 사용하여 학습
  - 기존 최적화 기반 모델에 비해 속도와 정확도를 향상
- **RAFT**(Teed, 2020) >> 최신 SOTA 달성 모델
  - 모든 픽셀 쌍의 상호작용을 반복적으로 계산하여 정확도를 극대화

- **기존 DL기반 연구의 한계**
  - 대부분의 딥러닝 기반 Optical Flow 모델은 합성 데이터셋을 사용하여 학습
  - 일반화 성능이 낮으며 domain shift 문제가 존재

2. Related Work

Self-Supervised Optical Flow

자가 지도 학습 Self-Supervised Optical Flow

- 라벨링이 되지 않은 원천데이터(비디오)에서 Optical Flow를 학습
- 재투영 오류 reprojection error를 최소화

- 관련 연구
  - UnFlow(Meister, 2018) : 쌍방향 일관성을 활용, 지도 없이 Optical Flow 학습
  - Uflow(Jonschkowski, 2020) : 학습 과정에서 중요한 픽셀을 강조, 최적화를 통한 성능 개선

- 한계
  - 대부분 동일한 도메인 내의 데이터에서 학습과 평가를 모두 진행
  - 다른 데이터에서의 일반화는 제한적

2. Related Work

Single Image Depth Estimation

단일 이미지 깊이 추정 Single Image Depth Estimation

: 단일 이미지에서 각 픽셀의 깊이 값을 예측하며, 기존 깊이 추정과 달리 단일 이미지만 사용

- 지도 학습 Supervised Learning
  - 대규모 깊이 데이터셋(ex KITTI) 필요
- Self-Supervised Learning
  - Stereo Depth Estimaiton : 다른 각도에서 찍힌 같은 장면을 활용, 두 이미지 사이의 disparity를 계산하여 깊이 추정
  - Monocular video based Depth Estimation : 단일 카메라로 촬영된 연속된 비디오 프레임을 활용, 예측된 깊이와 움직임
    다음 프레임을 복원하여 실제 이미지와 비교하여 오류를 최소화
- MiDaS(Ranftl, 2020)
  - 여러 데이터셋을 혼합하여 학습한 Depth Estimation 모델
  - 일반화 성능이 뛰어나며, 본 논문에서 제안한 Depthstillation에도 사용됨

2. Related Work

View Synthesis

뷰 생성 View Synthesis

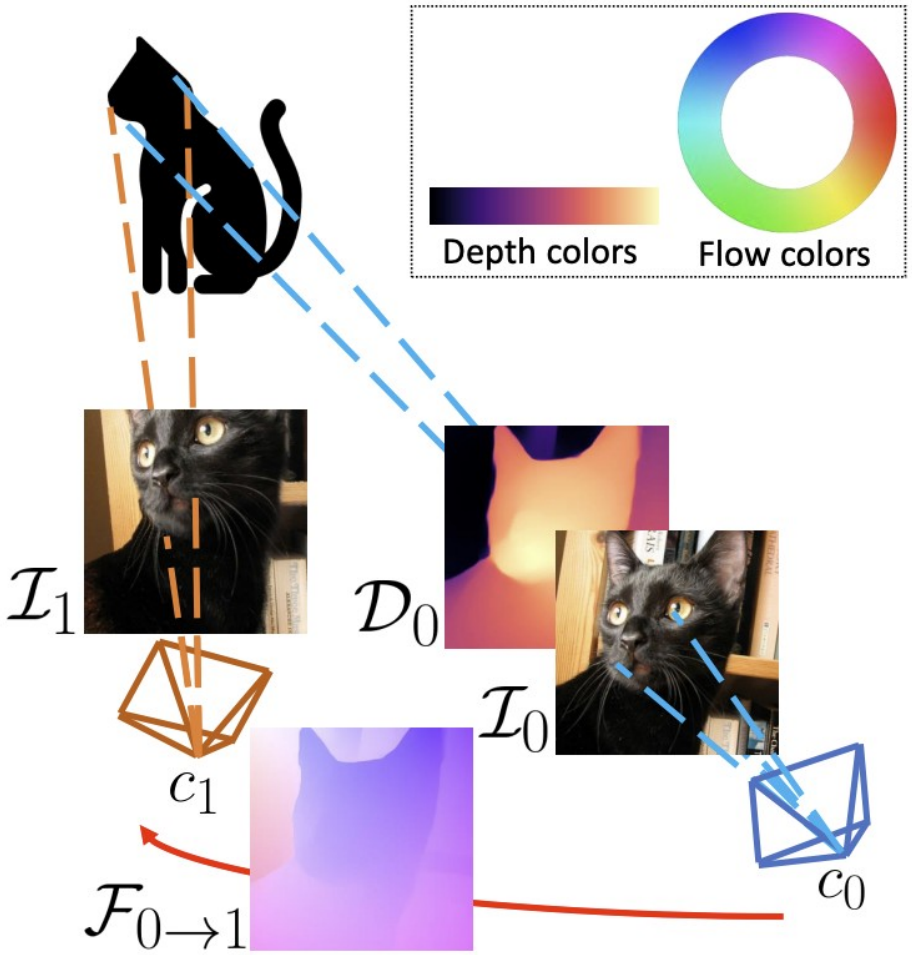
: 새로운 시점에서의 이미지를 생성하는 작업 >> Optical flow가 필수적

- 주요 응용
  - Video Interpolation : 두 프레임간 중간 프레임 생성(Super-SloMo, 2019)
  - 단일 이미지에서 3D효과(ken burns)생성

- 본 논문은 View Synthesis를 학습 데이터 생성에 사용

3. Method – Depthstillation

Depthstillation pipeline



1. 입력 이미지 I에서 깊이 맵 D 추정

$$D_0 = \Phi(I_0)$$

2. 가상 카메라 이동 Forward warping

$$p_1 \sim KT_{0\to1}D_0(p_0)K^{-1}p_0$$

3. Optical flow 계산

$$F_{0\to1} = p_1 - p_0$$

3. Method – Depthstillation

Depthstillation Pipeline -

Hole Filling

- Forward Warping 과정에서 일부 픽셀이 공백(holes)을 남길 수 있음
- 특히 물체의 경계 부분이나 가려진 영역에서 이러한 현상이 자주 발생

- Hole Mask : 공백이 발생한 픽셀 영역을 식별하기 위해 만들어지는 binary mask

$$P = (\mathcal{M}' == \mathcal{M}), \quad H' = H \cdot P$$

- Impainting을 통해 픽셀을 보충 >> 공백을 채우고 배경과 객체가 섞이는 문제를 줄임

3. Method – Depthstillation

Depthstillation Pipeline

3. Method – Depthstillation

Depthstillation Pipeline

Independent Motion

- 기존 Optical flow 기술은 대상이 움직이지 않고 카메라 시점만 변경되는, 정적인 환경에서만 동작
- 독립적으로 움직이는 객체를 다루지 못함

- 이를 해결하기 위해 Instance Segmentation을 사용해 객체를 분리, 각 객체에 대한 개별적인 움직임을 적용

$$\Pi = \{\Pi_i, i \in [1, N]\} \qquad \Pi = \Omega(I_0)$$

2) 객체 이동 시뮬레이션

$$p_1 \sim \begin{cases} KT_{0\to 1}D_0(p_0)K^{-1}p_0 \\ KT_{0\to \pi_i}D_0(p_0)K^{-1}p_0 \end{cases}$$

2025.01.08

III Experiments & Results

- Dataset
- EPE, F1-error
- **Results**

04. Experiments

Experiments – dataset

- **Train**
  - Depthstillation을 통해 생성된 가상 Optical Flow 데이터셋
    - **dCOCO**: COCO 데이터셋으로부터 생성된 Optical Flow 데이터
    - **dDAVIS**: DAVIS 데이터셋으로부터 생성된 Optical Flow 데이터
    - 기존 합성 데이터셋
      - FlyingChairs, FlyingThings3D

- **Test**
  - KITTI 2012, KITTI 2015 >> 실제 환경에서 촬영된 Optical Flow 데이터셋
  - 합성 데이터셋 MPI-Sintel

04. Experiments

Experiments – evalueation

평가 지표

- EPE(End-Point Error)
  - 예측된 optical flow와 ground truth간의 평균 유클리드 거리
- F1-Error
  - Ground Truth와 3픽셀 이상 차이나는 픽셀의 비율

- 기존 합성 데이터셋으로 학습된 RAFT, PWC-Net과 성능 비교

04. Results

Results

	Depth est.	Hole fill.	Moving obj.	Sintel C. EPE	Sintel C. > 3	Sintel F. EPE	Sintel F. > 3	KITTI 12 EPE	KITTI 12 Fl	KITTI 15 EPE	KITTI 15 Fl
(A)	✗	✗	✗	5.50	18.22	6.08	20.83	3.31	18.95	10.51	35.52
(B)	✓	✗	✗	2.52	7.17	3.72	11.04	2.02	7.53	4.84	16.26
(C)	✓	✓	✗	2.63	7.00	3.90	11.31	**1.82**	6.62	3.81	12.42
(D)	✓	✓	✓	**2.35**	**6.11**	**3.62**	**10.10**	1.83	**6.53**	**3.65**	**11.98**

Table 1. **Method ablation.** We train RAFT on dCOCO with different configurations of depthstillation: (A) constant depth for each image, (B) adding depth estimated by MiDaS [45], (C) adding hole-filling and (D) simulating object motions.

04. Results

Results

	Depth Model	Sintel C.		Sintel F.		KITTI12		KITTI15	
		EPE	> 3	EPE	> 3	EPE	Fl	EPE	Fl
(A)	No depth	5.50	18.22	6.08	20.83	3.31	18.95	10.51	35.52
(B)	Megadepth [26]	2.91	7.51	3.99	11.55	**1.81**	7.11	4.10	13.70
(C)	MiDaS [45]	**2.63**	**7.00**	**3.90**	**11.31**	1.82	**6.62**	**3.81**	**12.42**

Table 2. **Impact of depth estimator.** We train RAFT on dCOCO without depth estimation (A), using depth maps provided by MegaDepth (B) or MiDaS (C).

04. Results

Results

	# Training samples			Sintel C.		Sintel F.		KITTI12		KITTI15	
	Images	Motions	Total	EPE	> 3	EPE	> 3	EPE	Fl	EPE	Fl
(A)	4K	×1	4K	2.73	6.96	3.97	11.09	1.86	6.81	3.93	12.56
(B)	4K	×5	20K	2.56	6.78	3.88	10.99	**1.77**	**6.62**	3.93	12.57
(C)	20K	×1	20K	2.63	7.00	3.90	11.31	1.82	**6.62**	**3.81**	12.42
(D)	20K	×5	100K	**2.37**	**6.69**	**3.64**	**10.73**	1.79	6.79	3.82	**12.39**

Table 3. **Impact of images and virtual motions.** We train several RAFT models by changing the number of input images taken from COCO and the number of motions depthstilled for each one.

04. Results

Results

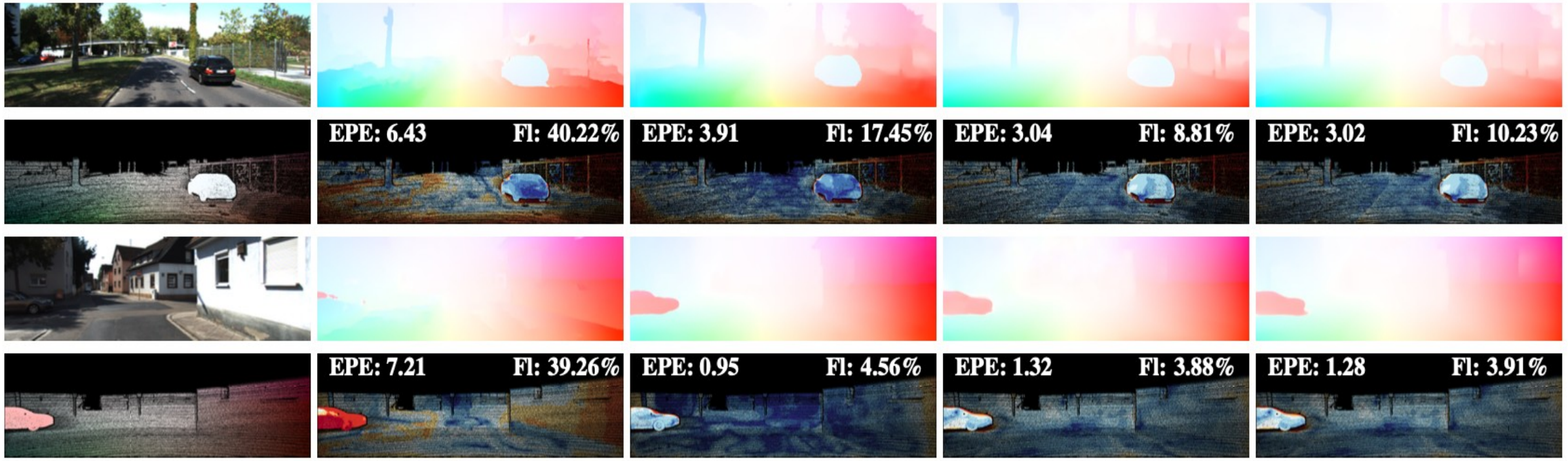


Figure 5. **Qualitative results on the KITTI 2015 training set.** On two rows: a) reference frame (top) and ground-truth flow (bottom), optical flow maps (top) by RAFT trained on b) Ch, c) Ch→Th, d) dCOCO and e) Ch→Th→dCOCO and error maps (bottom).

04. Results

Results

	Dataset	Sintel C.		Sintel F.		KITTI 12		KITTI 15	
		EPE	>3	EPE	>3	EPE	Fl	EPE	Fl
(A†)	Ch	2.26	7.35	4.51	12.36	4.66	30.54	9.84	37.56
(B†)	Ch→Th	1.46	4.40	2.79	8.10	2.15	9.30	5.00	17.44
(A)	Ch	2.36	7.70	4.39	12.04	5.14	34.64	10.77	41.08
(B)	Ch→Th	**1.64**	**4.71**	**2.83**	**8.67**	2.40	10.49	5.62	18.71
(C)	dCOCO	2.63	7.00	3.90	11.31	1.82	**6.62**	3.81	**12.42**
(D)	Ch→Th→dCOCO	1.88	5.31	3.23	9.26	**1.78**	7.00	**3.42**	13.08

Table 4. **Comparison with synthetic datasets – generalization.** Generalization achieved by RAFT when trained on synthetic data (A),(B), on our dCOCO dataset (C) and a combination of both (D). † are obtained with publicly available weights by [53] (2× GPUs).

04. Results

Results

	Pre-training	Fine-tuning	KITTI12		KITTI15	
			EPE	Fl	EPE	Fl
(A)	Ch	✗	5.14	34.64	15.56	47.29
	Ch	✓	1.42	4.86	2.40	8.49
(B)	Ch→Th	✗	2.40	10.49	9.04	25.53
	Ch→Th	✓	1.36	4.67	2.22	8.09
(C)	dCOCO	✗	1.82	6.62	5.09	16.72
	dCOCO	✓	1.37	4.70	2.76	9.15
(D)	Ch→Th→dCOCO	✗	1.78	7.00	4.82	18.03
	Ch→Th→dCOCO	✓	**1.32**	**4.54**	**2.21**	**7.93**

Table 5. **Comparison with synthetic datasets – fine-tuning.** Performance of RAFT variants pre-trained on synthetic datasets (A) and (B), on dCOCO (C) or both (D) when fine-tuned on a subset of 160 images from KITTI 2015, tested on KITTI 2012 and the remaining 40 images from KITTI 2015.

04. Results

Results

Model	Dataset	Sintel C.		Sintel F.		KITTI12		KITTI15	
		EPE	> 3	EPE	> 3	EPE	Fl	EPE	Fl
(A) PWCNet	Ch	3.33	-	4.59	-	5.14	28.67	13.20	41.79
(B) PWCNet	Ch→Th	**2.55**	-	**3.93**	-	4.14	21.38	10.35	33.67
(C) PWCNet	dCOCO	4.14	11.54	5.57	15.58	**3.16**	**13.30**	**8.49**	**26.06**
(D) RAFT	dCOCO	2.63	7.00	3.90	11.31	1.82	6.62	3.81	12.42

Table 6. **Impact of depthstillation on different architectures.** Evaluation on PWCNet and RAFT. Entries with ”-” are not provided in the original paper.

04. Results

Results

	Model	Dataset	KITTI12		KITTI15	
			EPE	Fl	EPE	Fl
(A)	UFlow	DAVIS	3.49	14.54	9.52	25.52
(B)	PWCNet	dDAVIS	**2.81**	**11.29**	**6.88**	**21.87**
(C)	RAFT	dDAVIS	1.78	6.85	3.80	13.22

Table 7. **Comparison between self-supervision and depthstillation – generalization.** Effectiveness of the two strategies when evaluated on unseen data (KITTI 2012 and 2015).

04. Results

Results

	Model	Dataset	KITTI12		KITTI15	
			EPE	Fl	EPE	Fl
(A)	UFlow	KITTI	-	-	**3.08**	**10.00**
(B)	PWCNet	dKITTI	2.64	9.43	7.92	22.17
(C)	RAFT	dKITTI	1.76	5.91	4.01	13.35

Table 8. **Comparison between self-supervision and depthstillation – specialization.** Effectiveness of the two strategies when training and testing on similar data (KITTI 2015). Entries with ”-” are not provided in the original paper.

05. Conclusion

Conclusion

- Depthstillation을 활용하여 기존 합성 데이터보다 더 양질의 데이터를 효율적으로 얻어낼 수 있었음
  - 타 데이터셋에서의 일반화 성능 또한 뛰어나기 때문에 실제 환경에서의 사용이 용이
  - 단일 이미지와 가상 카메라 움직임만으로 대규모 학습데이터 생성이 가능

- 한계점
  - 근본적으로는 실제 데이터를 기반으로 하지 않기 때문에 복잡한 장면에서 일부 제한이 있을 수 있음
  - 향후 연구에서 더 다양한 실제 시나리오를 활용할 필요가 있음