Perspective Flow Aggregation for Data-Limited 6D Object Pose Estimation

Yinlin Hu [1,2,3], Pascal Fua [1], Mathieu Salzmann [1,2]

[1] EPFL Computer Vision Lab, [2] ClearSpace, [3] Magic Leap

Problems

- Most recent 6D object pose estimation methods, including unsupervised ones, require many real training images.
- Unfortunately, for some applications, such as those in space or deep under water, acquiring real images, even unannotated, is virtually impossible. These are the scenarios we refer to as **data-limited**.
- Although rendering-based synthetic techniques can help, the domain shift between the synthetic and real images is still a problem.

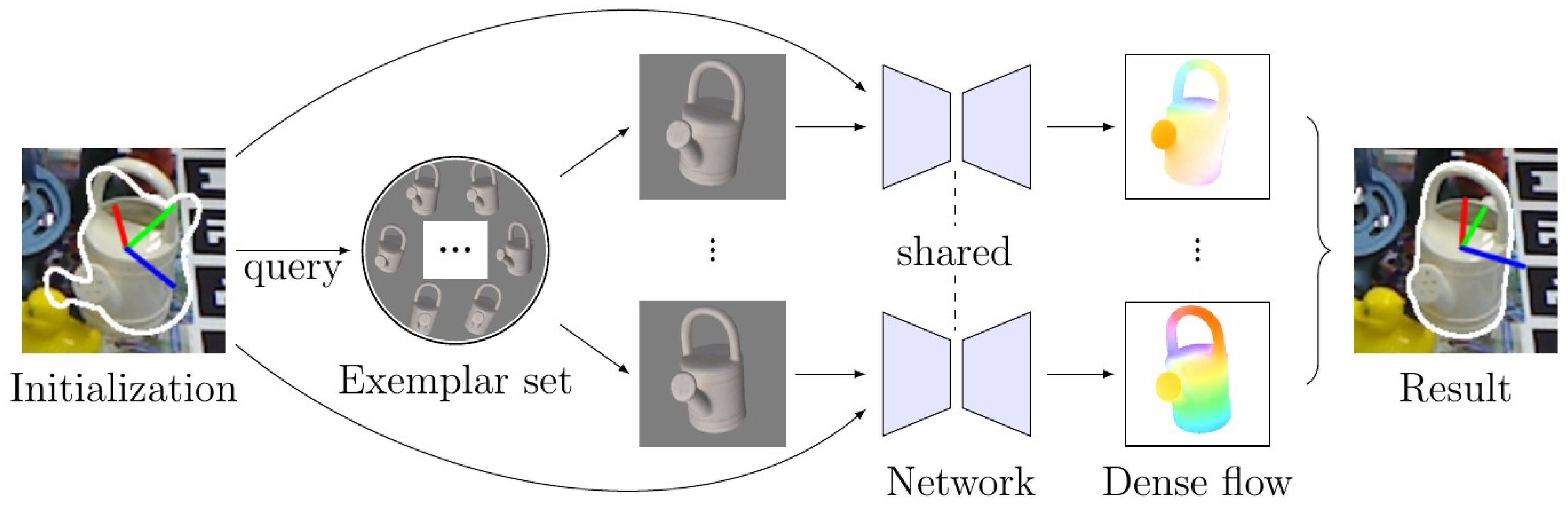
(a) (b) (c) (d) (e)

(a) Synthesized images. **(b)** Although the resulting accuracy on synthetic data is great, **(c)** that on real images is significantly worse. **(d)** While the common global-based refinement approach can help, it still suffers from the synthetic-to-real domain gap (DeepIM). **(e)** Our local-based strategy generalizes much better to real images.

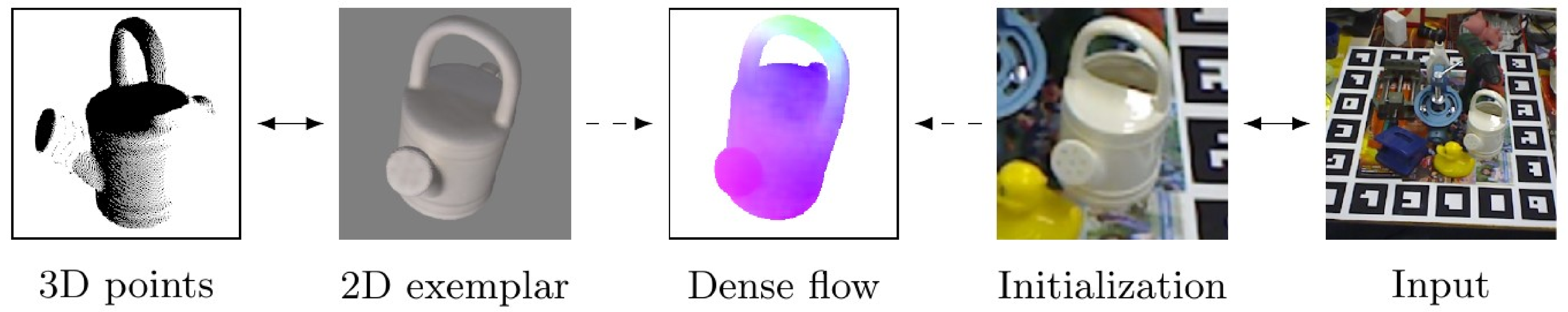
Code is available at: https://github.com/cvlab-epfl/perspective-flow-aggregation

Solution

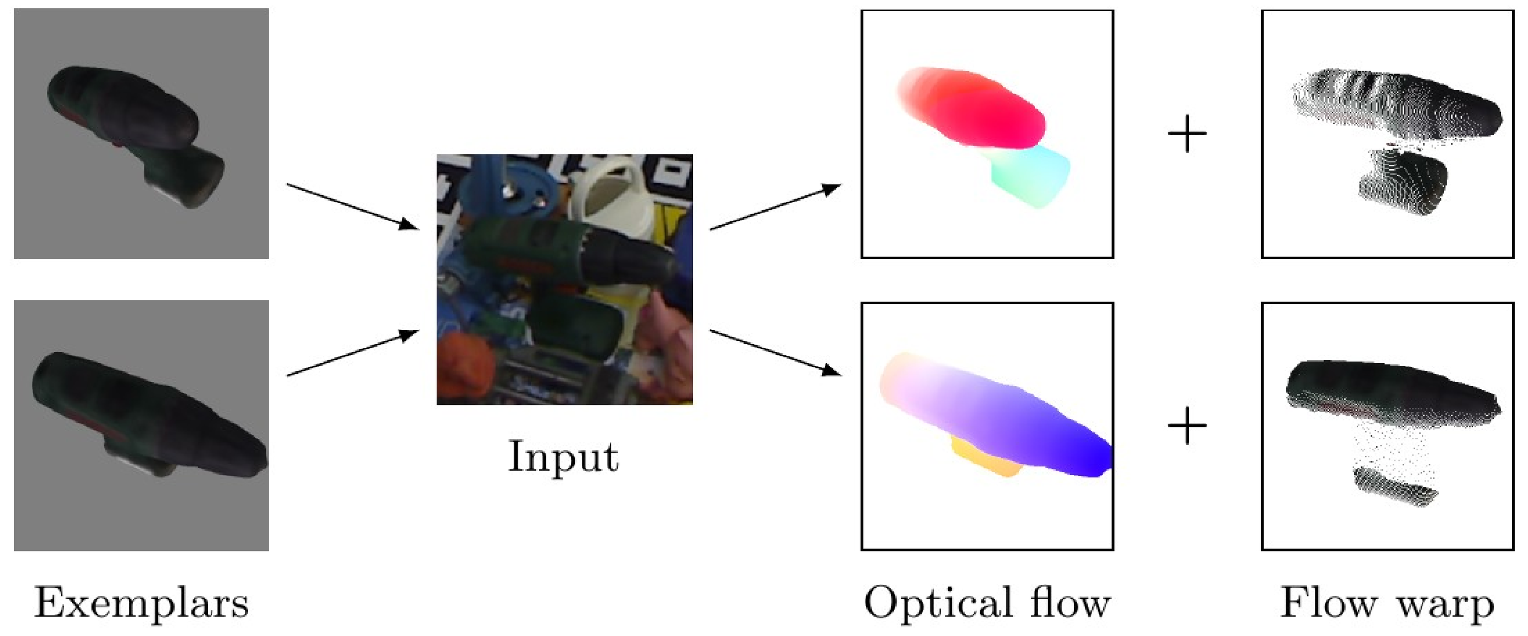
- We propose estimating dense 2D-to-2D local correspondences between input images to force the supervision of our training to occur at the pixel-level, making our DNN learn to extract features that contain lower-level information and thus generalize across domains:



- From optical flow to pose refinement:

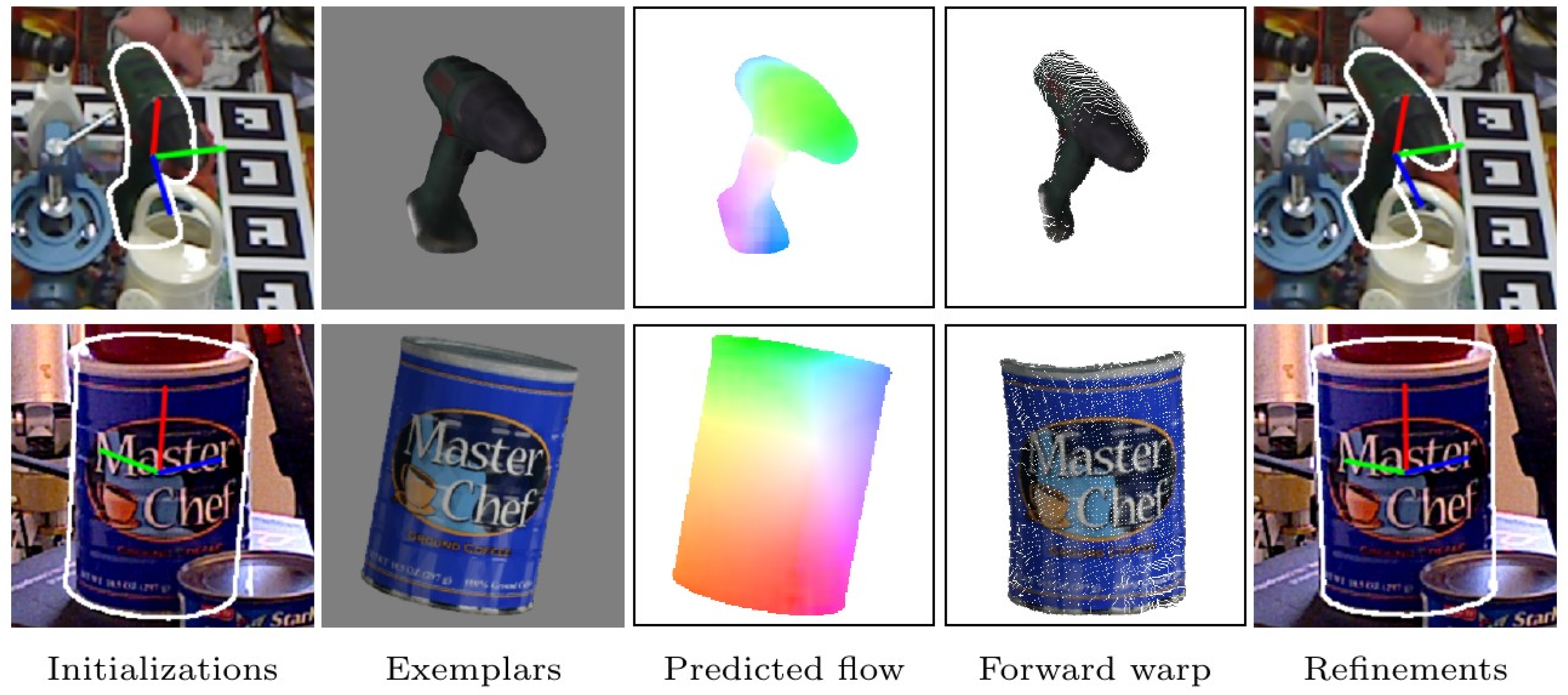


- Multi-view flow aggregation:



Experiments

Cross-domain refinement (model is trained purely on synthetic data):



Even more accurate than annotations sometimes:

Comparing against the SoTA:

Data	Metrics	PoseCNN	SegDriven	PVNet	GDR-Net	DeepIM	CosyPose	**Ours (+0)**	**Ours (+20)**
LM	ADD-0.1d	62.7	-	86.3	93.7	88.6	-	84.5	**94.4**
OLM	ADD-0.1d	24.9	27.0	40.8	62.2	55.5	-	48.2	**64.1**
YCB	ADD-0.1d	21.3	39.0	-	60.1	-	-	56.4	**62.8**
	AUC	61.3	-	73.4	84.4	81.9	84.5	76.8	**84.9**

Training with some real images on OLM:

	0	10	20	90	180
DeepIM	41.1	45.6	48.2	58.1	61.4
CosyPose	42.4	46.8	48.9	58.8	61.9
Ours	**48.2**	**59.5**	**64.1**	**64.9**	**65.3**

Multi-view flow aggregation:

Initialization		N=1	N=2	N=4	N=8
NS	54.1	82.0	82.7	**84.3**	84.1
HSV	52.7	81.1	81.2	83.3	**83.9**
NS+HSV	60.2	82.0	83.4	84.5	**84.9**
FPS		~32	~25	~20	~14