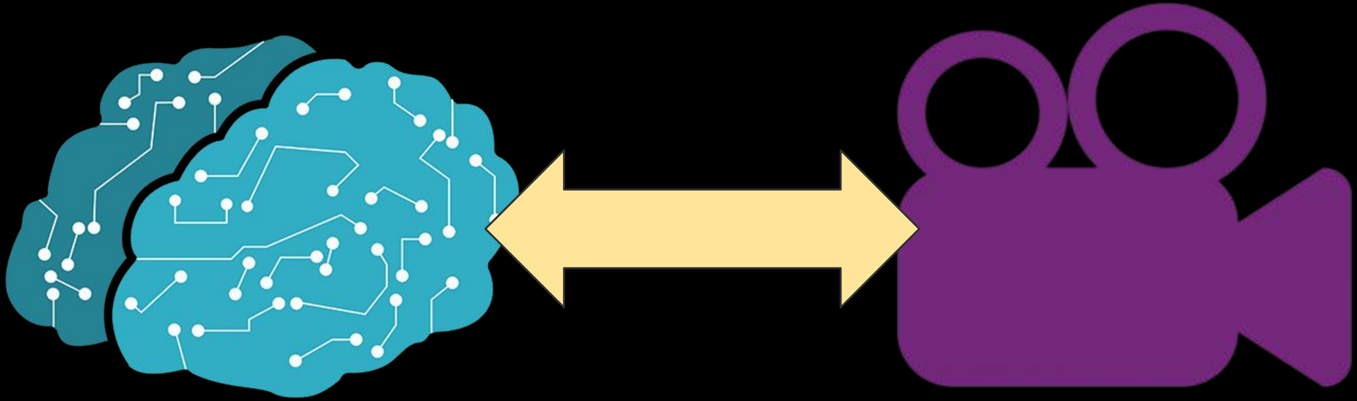
Multigrid Predictive Filter Flow for Unsupervised Learning on Videos



Shu Kong

Dept. of Computer Science

University of California, Irvine

Notice

Slides for academic use purpose; [link] for Google slides with videos.

The original slides contain multiple projects for in-person presentation.
This public version only includes a subset that is about mgPFF.

More contents will be released later this year when no conflicts are involved; or provided with personal request.

Regards,
Shu

Video

Video

Video is an electronic medium for the recording, copying, playback, broadcasting, and display of moving visual media.

Video

Video is an electronic medium for the recording, copying, playback, broadcasting, and display of moving visual media.

transporting information, entertainment, surveillance, assistive robots, autonomous vehicle...

Video

Video is an electronic medium for the recording, copying, playback, broadcasting, and display of moving visual media.

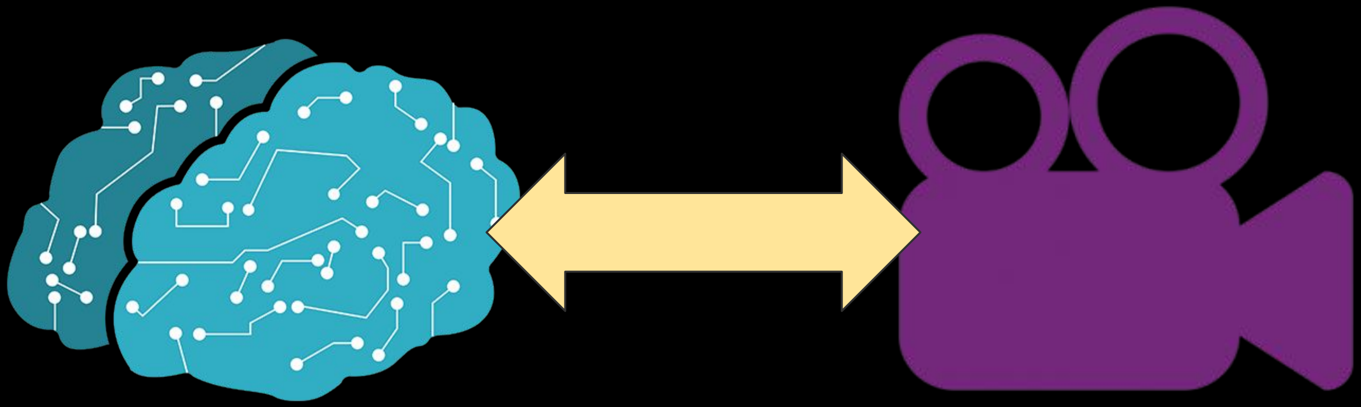
transporting information, entertainment, surveillance, assistive robots, autonomous vehicle...

natural signal, free and massive in amount...
often coming with free audio/caption.

Video Mining

To train machines using videos --

- What to use from the videos?
- How to train? Train for what?
- What/how to make a difference?



Outline of Video Mining

1. Unsupervised Learning with Multigrid Predictive Filter Flow

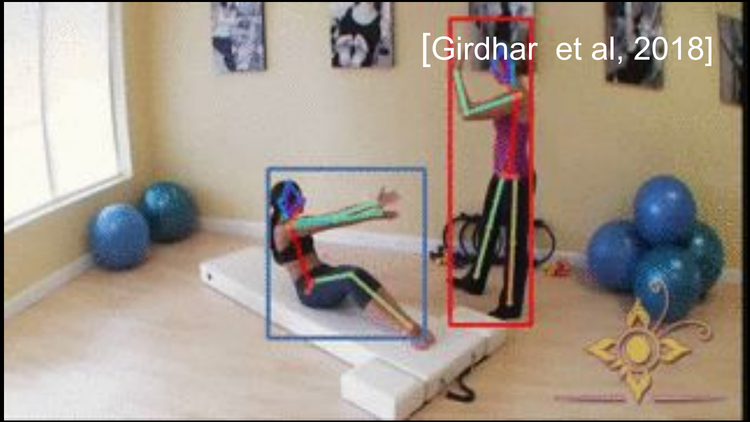
   video inst. seg./tracking, pose tracking, long-range flow, video shot det.

[Kong & Fowlkes, 2019]

Fully-Supervised Learning for Tracking

fully-supervised learning for tracking

pro: state-of-the-art tracking performance on benchmarks

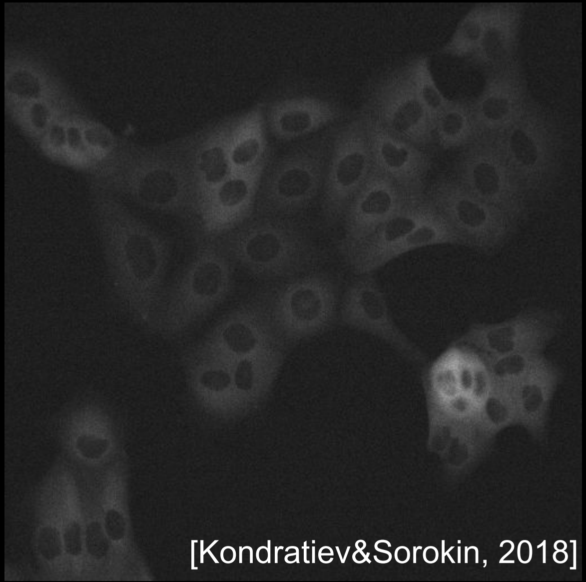


Fully-Supervised Learning for Tracking

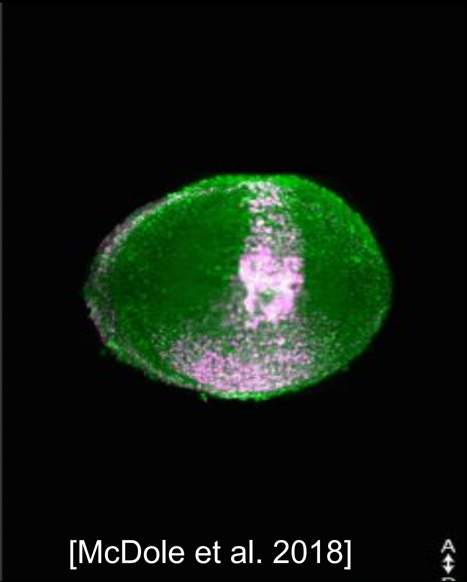
fully-supervised learning for tracking

pro: state-of-the-art tracking performance on benchmarks

con: expensive to annotate training set, domain/data bias



[Kondratiev&Sorokin, 2018]



[McDole et al. 2018]

Unsupervised Learning for Tracking

pro: w/o annotation, domain-agnostic,

cognitively, 2-week newborn can track w/o knowing semantics

Principles of the Idea

- low-level vision based method;
  - better generalization, compact model, cognitive observation
- exploiting temporal consistency (frame reconstruction);
- allowing for learning transferable features, or on specific data;
- broader application.

Multigrid *Predictive* *Filter Flow* (mgPFF)

Method:

- making direct, fine-grained predictions of how to reconstruct a video frame from pixels of another frame
- being trained using simple photometric reconstruction error

Multigrid *Predictive* *Filter Flow* (mgPFF)

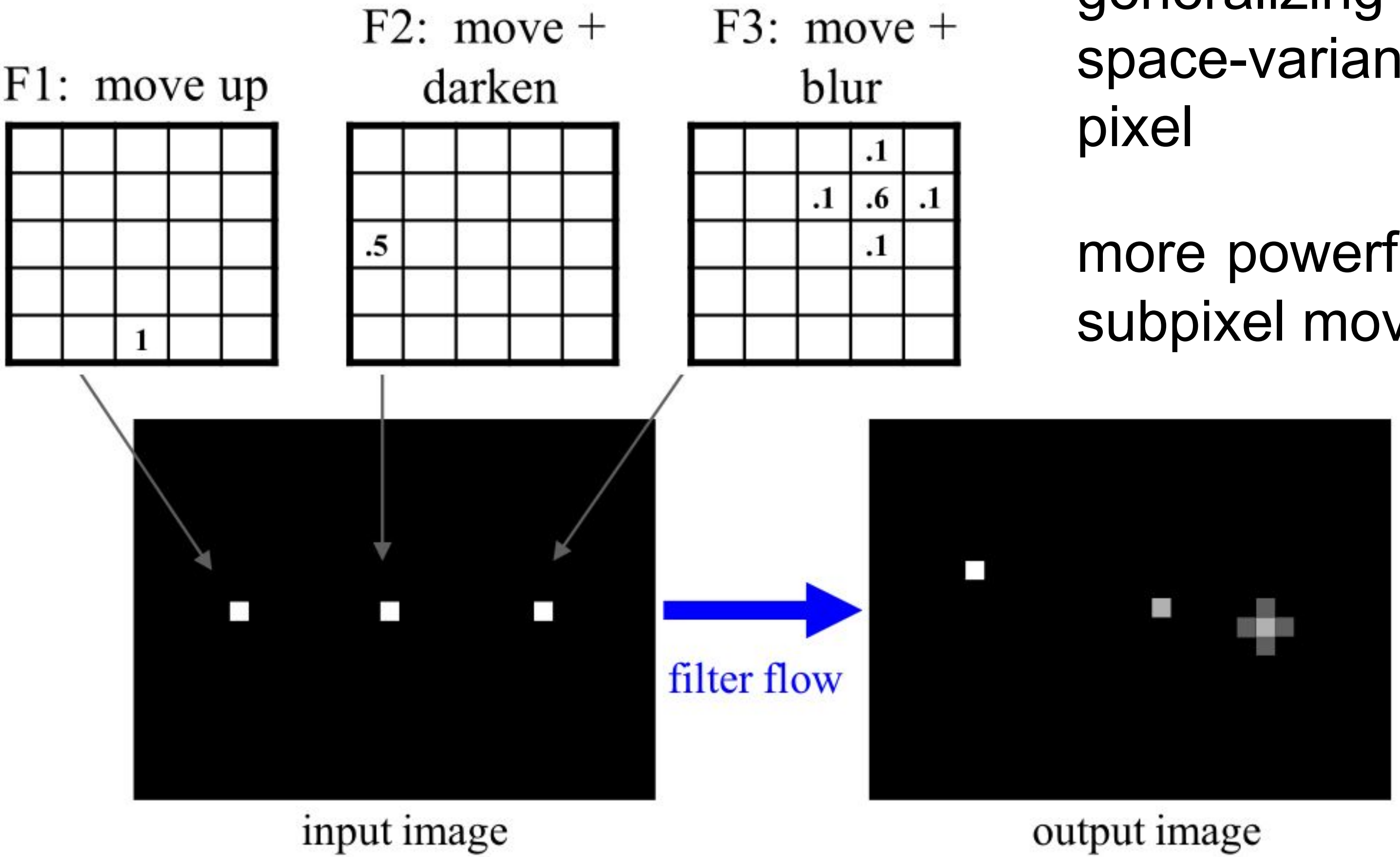
Method:

- making direct, fine-grained predictions of how to reconstruct a video frame from pixels of another frame
- being trained using simple photometric reconstruction error

Highlights:

- unsupervised learning on free-form videos with single GPU;
- easy training, long-range pixel connection;
- extremely compact, 4.6MB;
- fast computation, (0.1sec for a pair of 256x256 images).

Filter Flow [Seitz & Baker, 2009]



generalizing optical flow with space-variant filters at each pixel

more powerful in modeling subpixel movement

Filter Flow [Seitz & Baker, 2009]

models image transformation $\mathbf{I}_B \rightarrow \mathbf{I}_A$ by linear mapping, where each pixel in $\mathbf{I}_A$ only depends on local neighborhood centered at the same place in $\mathbf{I}_B$

$$\min_{T} \quad ||TI_1 - I_2||_2^2$$

$$\mathbf{I}_A = \mathbf{T}_{B\rightarrow A} \cdot \mathbf{I}_B$$

Filter Flow [Seitz & Baker, 2009]

pro: powerful, elegant, interpretable, applicable

stereo, optical flow, deblur, deconvolution, morphing, defocus, affine alignment

$$\mathbf{I}_A = \mathbf{T}_{B \rightarrow A} \cdot \mathbf{I}_B$$

Filter Flow [Seitz & Baker, 2009]

pro: powerful, elegant, interpretable, applicable

stereo, optical flow, deblur, deconvolution, morphing, defocus, affine alignment

con: optimization solver, impractical

e.g., 22 hrs for optical flow on an image pair of 584x388 resolution

$$\mathbf{I}_A = \mathbf{T}_{B\rightarrow A} \cdot \mathbf{I}_B$$

Filter Flow [Seitz & Baker, 2009]

pro: powerful, elegant, interpretable, applicable

stereo, optical flow, deblur, deconvolution, morphing, defocus, affine alignment

con: optimization solver, impractical

e.g., 22 hrs for optical flow on an image pair of 584x388 resolution

too slow☹

$$\mathbf{I}_A = \mathbf{T}_{B\rightarrow A} \cdot \mathbf{I}_B$$

Predictive Filter Flow [Kong & Fowlkes, 2018]

~~optimization-based solver~~ $\min_{T} \quad ||TI_1 - I_2||_2^2$

Predictive Filter Flow [Kong & Fowlkes, 2018]

~~optimization-based~~ solver $\min_{T} \quad ||TI_1 - I_2||_2^2$

we train a function/model $f_{\mathbf{w}}(\cdot)$ that predicts the transformation $\hat{\mathbf{T}}$ specific to image pair $(\mathrm{I}_1, \mathrm{I}_2)$ under the assumption that the image pairs $(\mathrm{I}_1, \mathrm{I}_2)$ are drawn from some fixed joint distribution

Predictive Filter Flow [Kong & Fowlkes, 2018]

~~optimization-based solver~~ $\min_{T} \quad ||TI_1 - I_2||_2^2$

we train a function/model $f_{\mathbf{w}}(\cdot)$ that predicts the transformation $\hat{\mathbf{T}}$ specific to image pair (I1, I2) under the assumption that the image pairs (I1, I2) are drawn from some fixed joint distribution

$$\mathbf{I}_2 \approx \hat{\mathbf{T}}\mathbf{I}_1, \quad \hat{\mathbf{T}} \equiv f_{\mathbf{w}}(\mathbf{I}_1, \mathbf{I}_2)$$

Predictive Filter Flow [Kong & Fowlkes, 2018]

$$\min_{\mathbf{w}} \sum_{i=1}^{N} \ell(\mathbf{I}_2^i - \hat{\mathbf{T}} \cdot \mathbf{I}_1^i) + \mathcal{R}(\hat{\mathbf{T}}),$$

$$\text{s.t. constraint on } \mathbf{w} \quad \hat{\mathbf{T}} \equiv f_{\mathbf{w}}(\mathbf{I}_1, \mathbf{I}_2)$$

With sampled image pairs $\{(\mathbf{I}_1^i, \mathbf{I}_2^i)\}$ we seek parameters $\mathbf{w}$ that minimize the difference between a recovered image $\hat{\mathbf{I}}_2$ and the real one $\mathbf{I}_2$ measured by some loss ℓ

Predictive Filter Flow [Kong & Fowlkes, 2018]

$$\min_{\mathbf{w}} \sum_{i=1}^{N} \ell(\mathbf{I}_2^i - \hat{\mathbf{T}} \cdot \mathbf{I}_1^i) + \mathcal{R}(\hat{\mathbf{T}}),$$

$$\text{s.t. constraint on } \mathbf{w} \quad \hat{\mathbf{T}} \equiv f_{\mathbf{w}}(\mathbf{I}_1, \mathbf{I}_2)$$

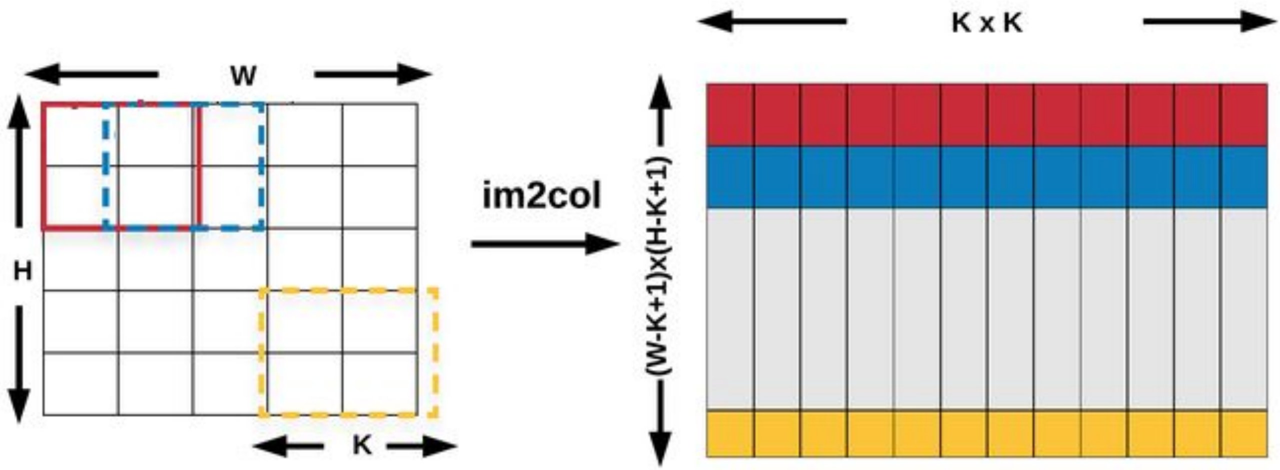
locality	Kernel size, im2col with inner product
$f_{\mathbf{w}}(\cdot)$	CNN
non-negativity	ReLU
sum-to-one	softmax

Predictive Filter Flow [Kong & Fowlkes, 2018]

$$\min_{\mathbf{w}} \sum_{i=1}^{N} \ell(\mathbf{I}_2^i - \hat{\mathbf{T}} \cdot \mathbf{I}_1^i) + \mathcal{R}(\hat{\mathbf{T}}),$$

$$\text{s.t. constraint on } \mathbf{w} \quad \hat{\mathbf{T}} \equiv f_{\mathbf{w}}(\mathbf{I}_1, \mathbf{I}_2)$$

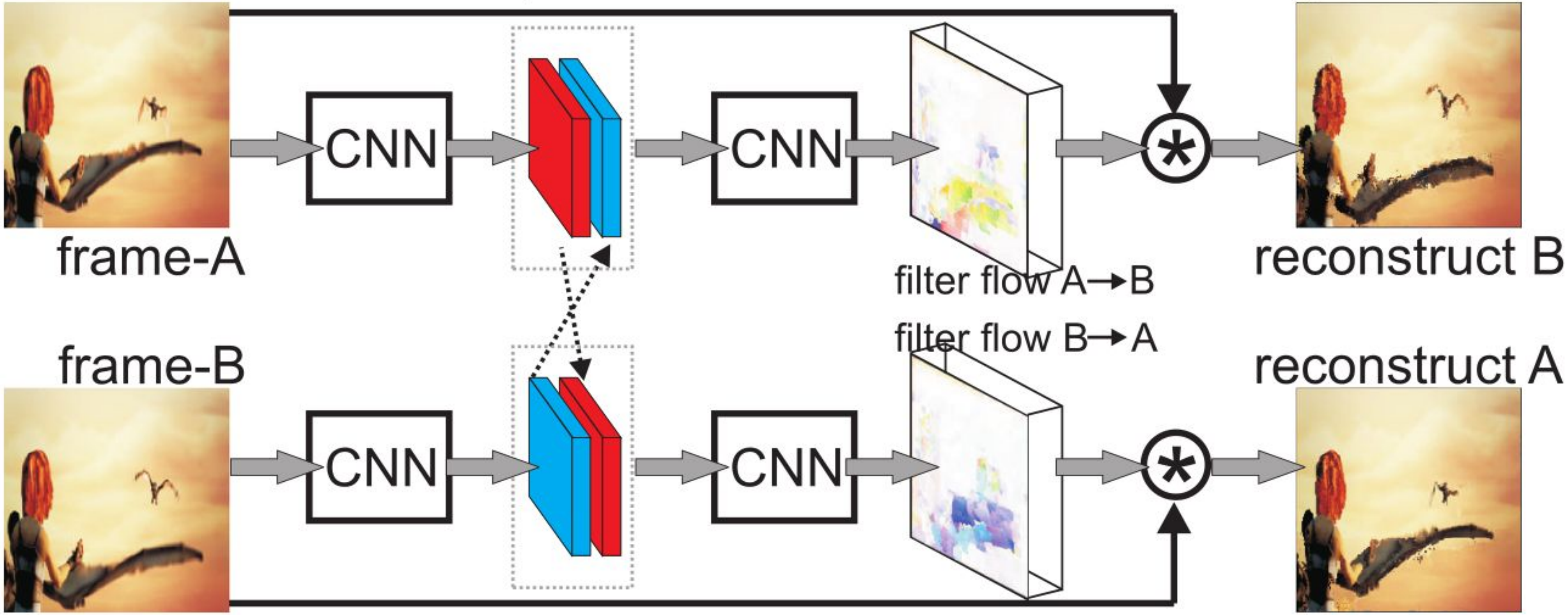
locality	Kernel size, im2col with inner product
$f_{\mathbf{w}}(\cdot)$	CNN
non-negativity	ReLU
sum-to-one	softmax



PFF for Unsupervised Learning on Videos

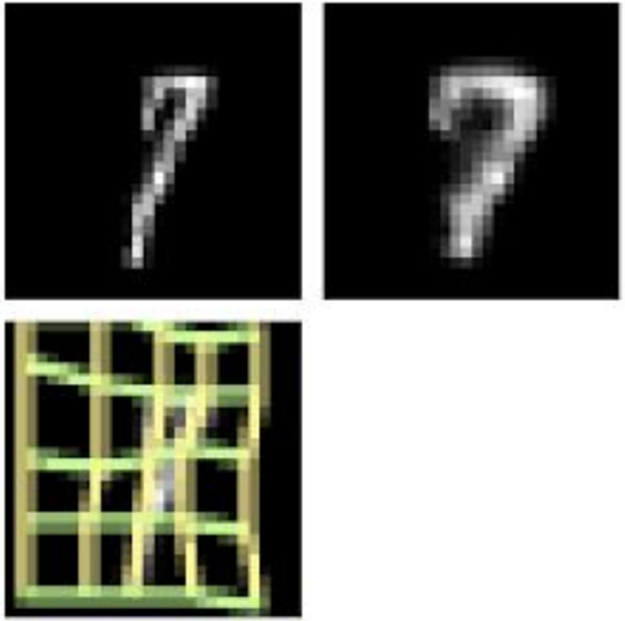
$$\min_{\mathbf{w}} \sum_{i=1}^{N} \ell(\mathbf{I}_2^i - f_{\mathbf{w}}(\mathbf{I}_1^i, \mathbf{I}_2^i) \cdot \mathbf{I}_1^i) + \mathcal{R}(f_{\mathbf{w}}(\mathbf{I}_1^i, \mathbf{I}_2^i))$$

constraints on the filter flow, losses between reconstruction and original frames



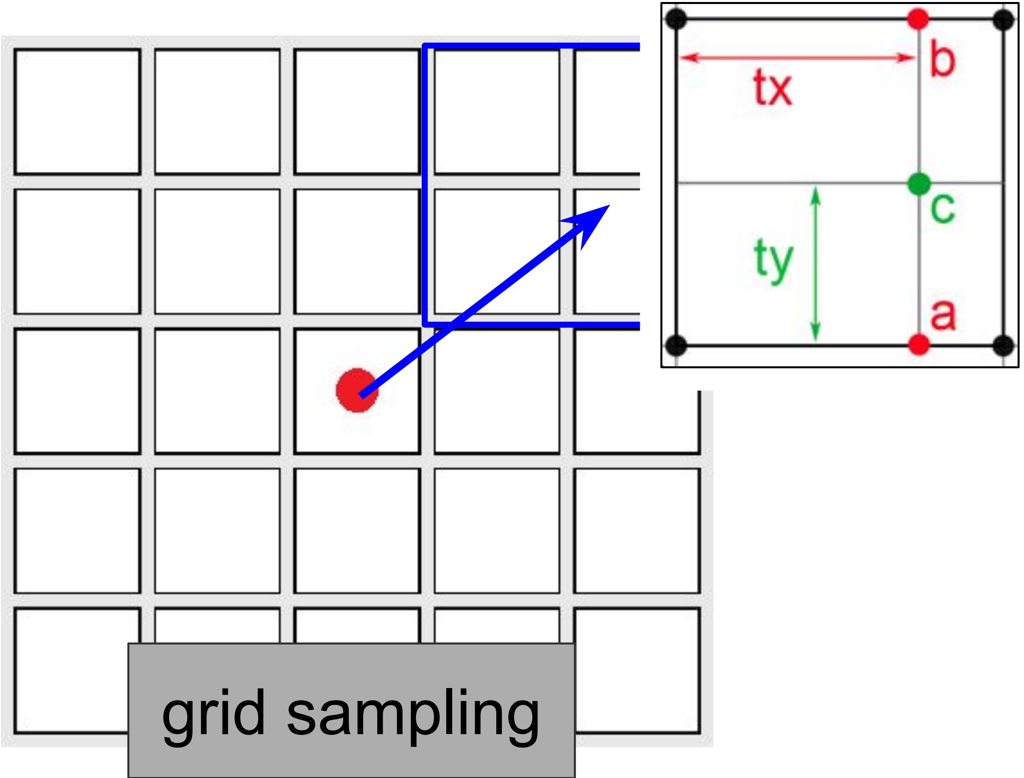
PFF for Unsupervised Learning on Videos

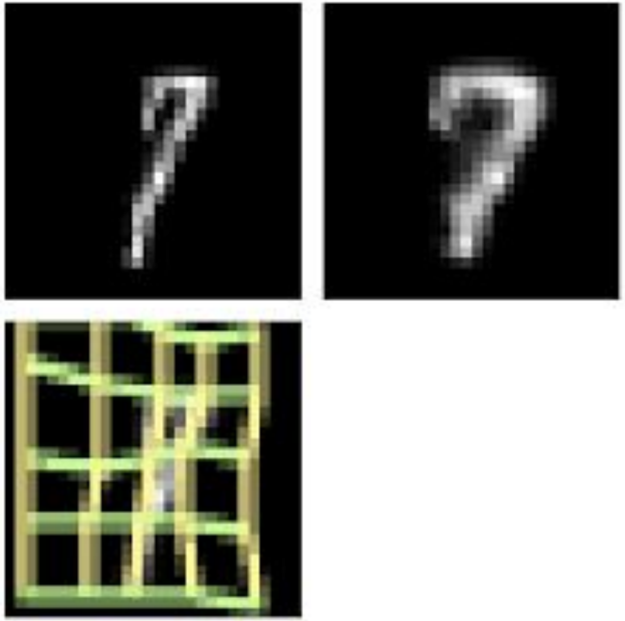
alternative: grid sampling layer in the Spatial Transformer Network [Jaderberg et al. 2015]



PFF for Unsupervised Learning on Videos

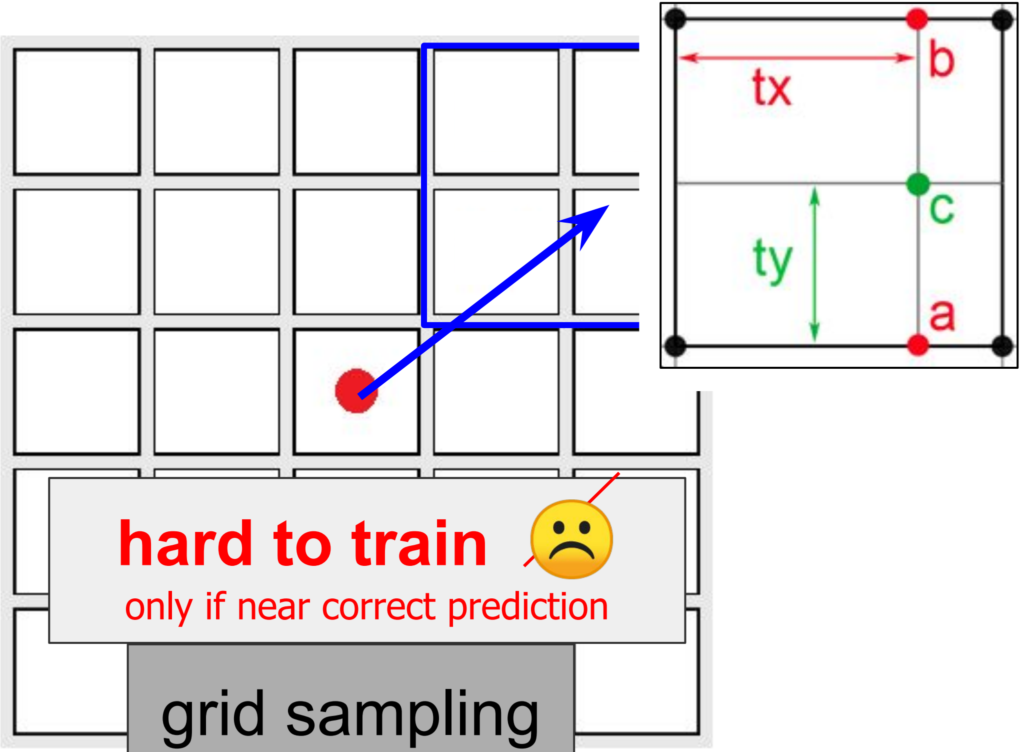
alternative: grid sampling layer in the Spatial Transformer Network [Jaderberg et al. 2015]





PFF for Unsupervised Learning on Videos

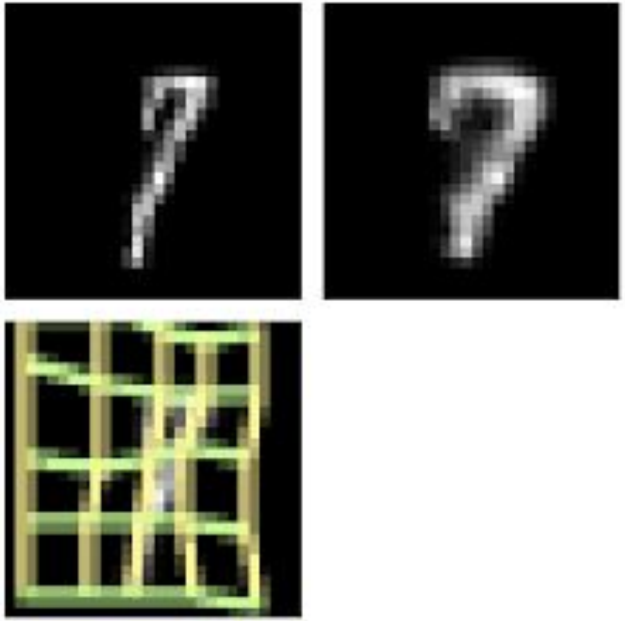
alternative: grid sampling layer in the Spatial Transformer Network [Jaderberg et al. 2015]



hard to train ☹

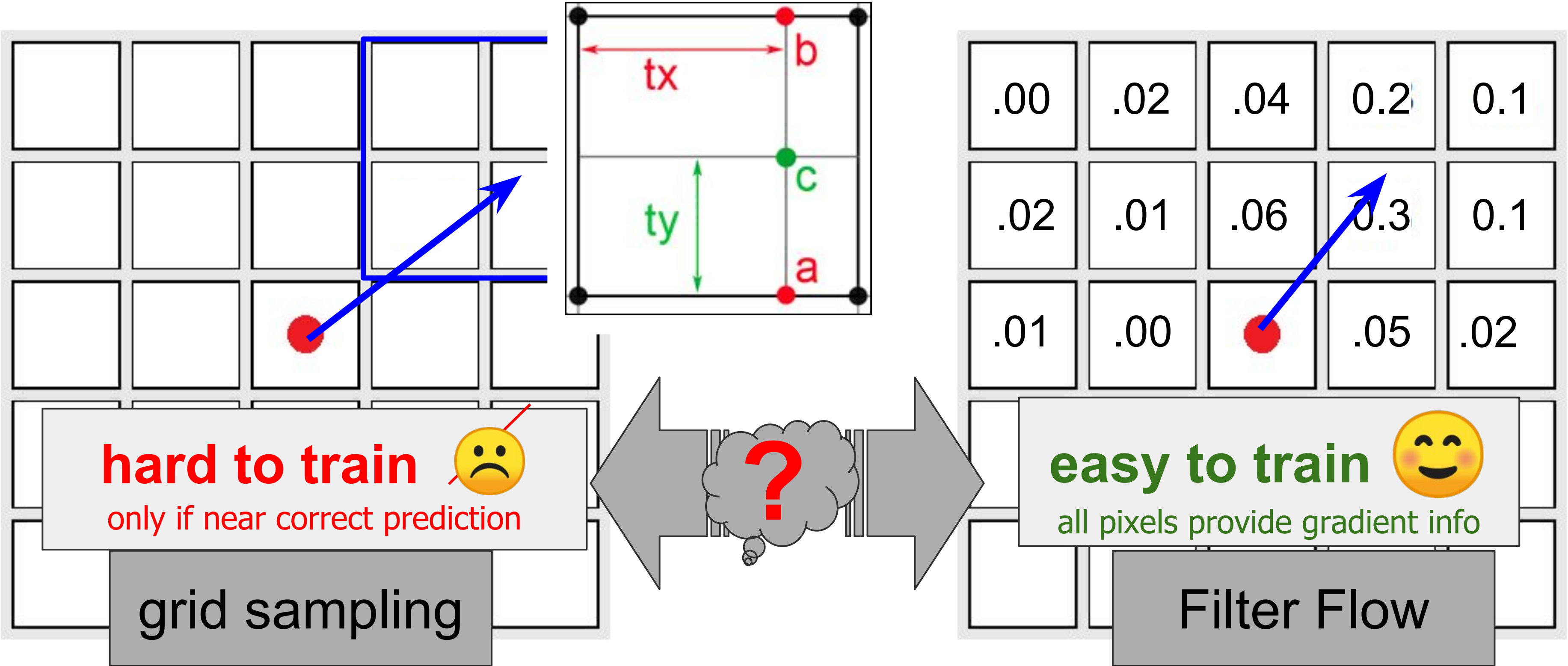
only if near correct prediction

grid sampling



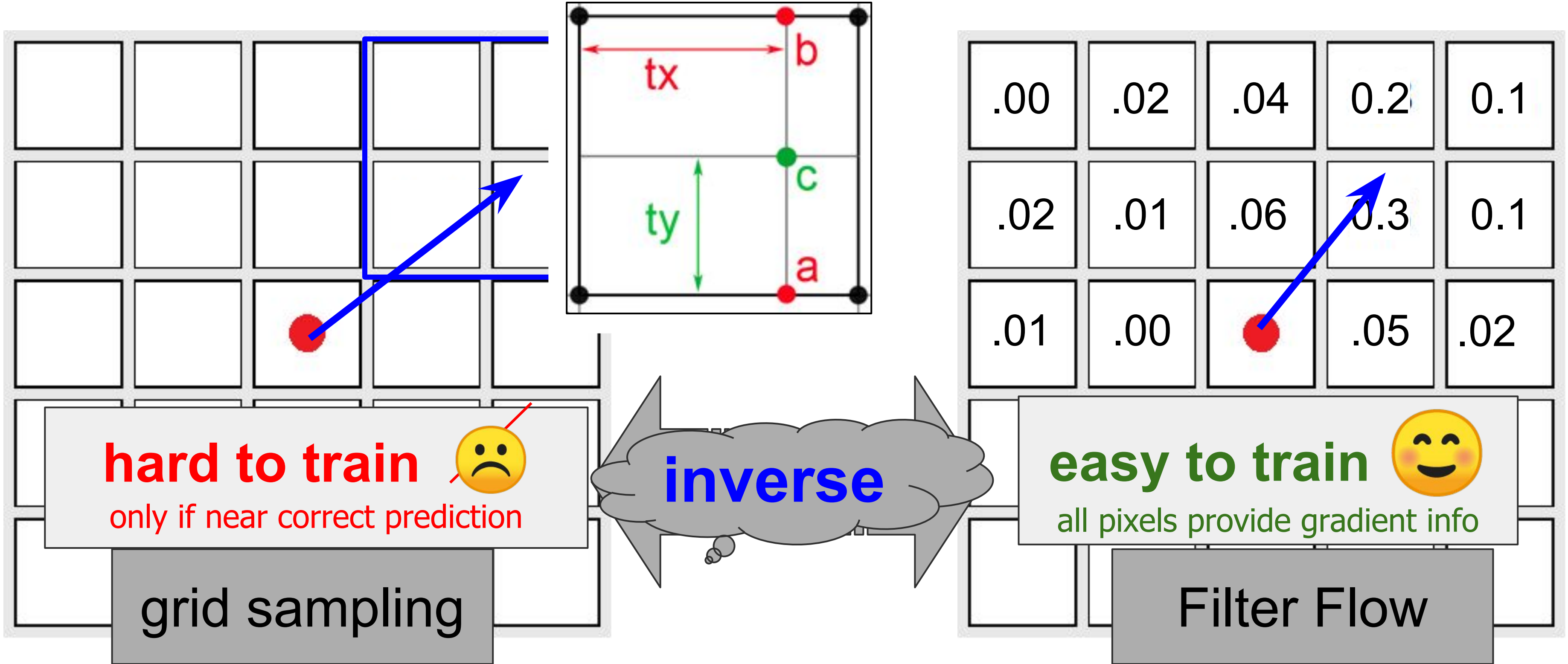
PFF for Unsupervised Learning on Videos

alternative: grid sampling layer in the Spatial Transformer Network [Jaderberg et al. 2015]



PFF for Unsupervised Learning on Videos

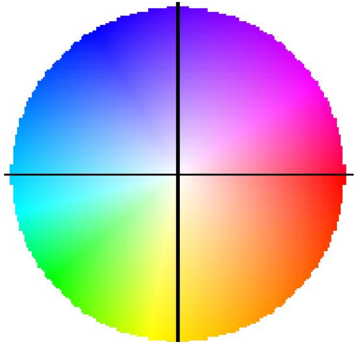
alternative: grid sampling layer in the Spatial Transformer Network [Jaderberg et al. 2015]

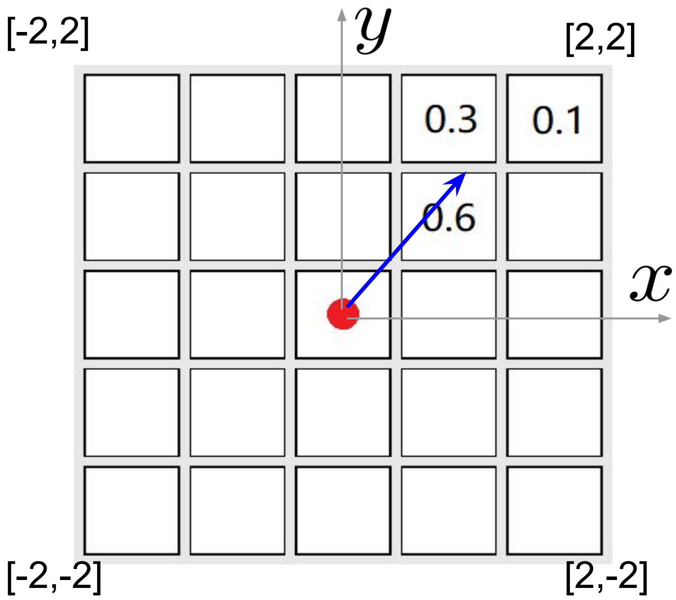


Voting for Offset

encouraging unimodal shape of the filter for the offset prediction

also allowing for visualization





$$\begin{bmatrix} v_x(i,j) \\ v_y(i,j) \end{bmatrix} = \sum_{x,y} \hat{T}_{ij,xy} \begin{bmatrix} x-i \\ y-j \end{bmatrix}$$

$$\begin{pmatrix} 1.1 \\ 1.4 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} *0.6 + \begin{pmatrix} 1 \\ 2 \end{pmatrix} *0.3 + \begin{pmatrix} 2 \\ 2 \end{pmatrix} *0.1$$

PFF for Unsupervised Learning on Videos

another challenge:

requiring large flow size to capture large displacement.

PFF for Unsupervised Learning on Videos

another challenge:

requiring large flow size to capture large displacement.

If pixel movement is in [-40, 40] pixels, then a filter flow size should be no less than 80, meaning 80x80 kernel for each pixel.

PFF for Unsupervised Learning on Videos

another challenge:

requiring large flow size to capture large displacement.

If pixel movement is in [-40, 40] pixels, then a filter flow size should be no less than 80, meaning 80x80 kernel for each pixel.

If the image is 256x256, then the output is 256x256x6400!

PFF for Unsupervised Learning on Videos

another challenge:

requiring large flow size to capture large displacement.

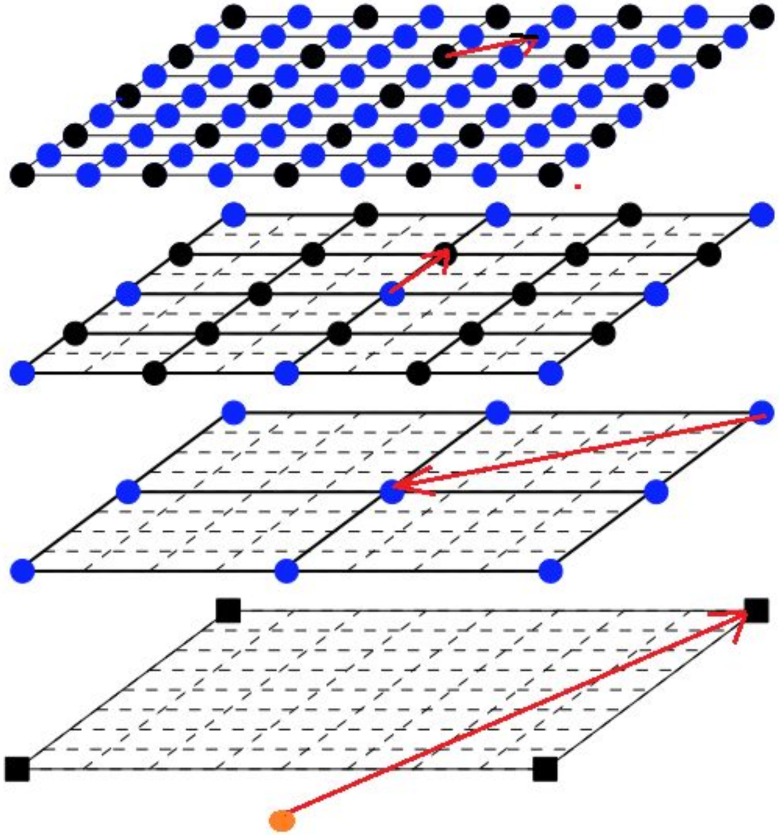
If pixel movement is in [-40, 40] pixels, then a filter flow size should be no less than 80, meaning 80x80 kernel for each pixel.

If the image is 256x256, then the output is 256x256x6400!

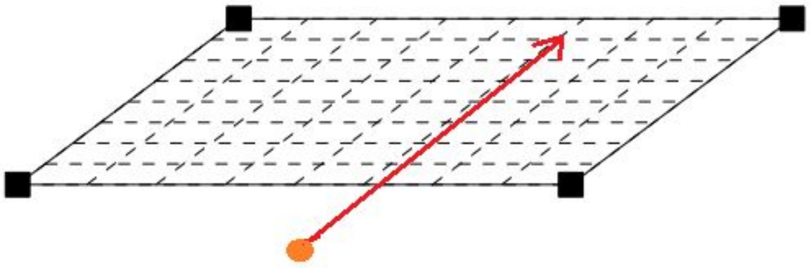
Our solution is Multigrid PFF.

Multigrid PFF for Large Displacement

multigrid filter

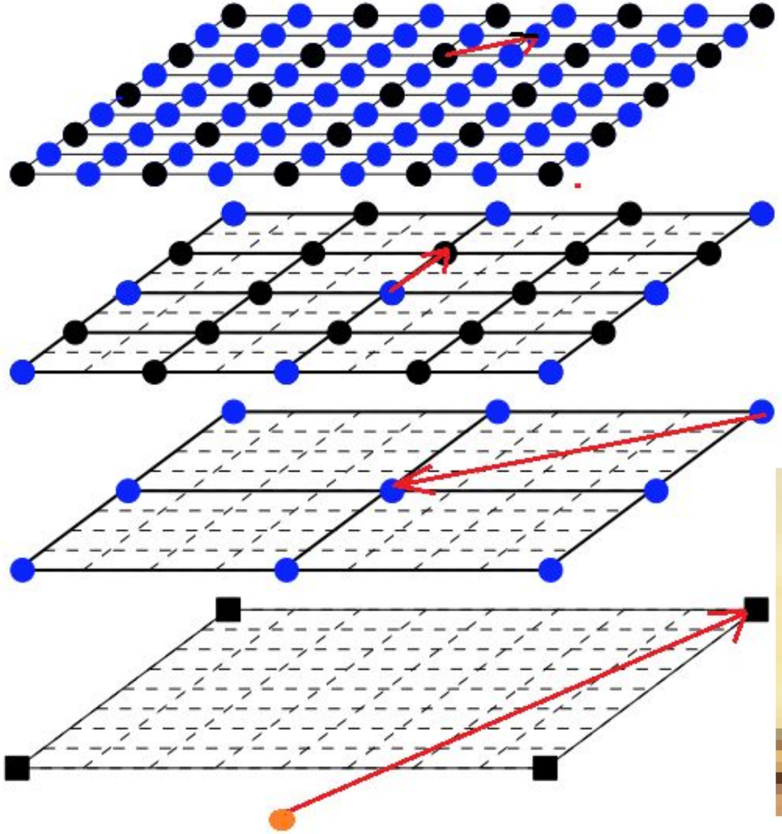


Decompose large sparse linear operator into a product of more compact terms



Multigrid PFF for Large Displacement

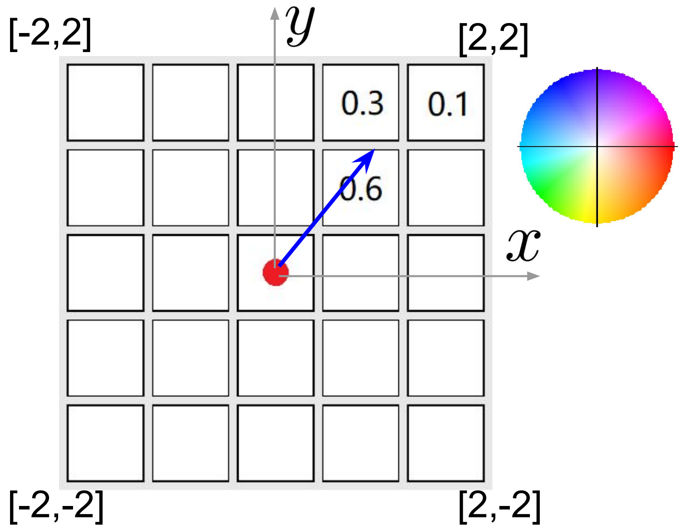
multigrid filter ⟷ coarse-to-fine



multi-resolution frames

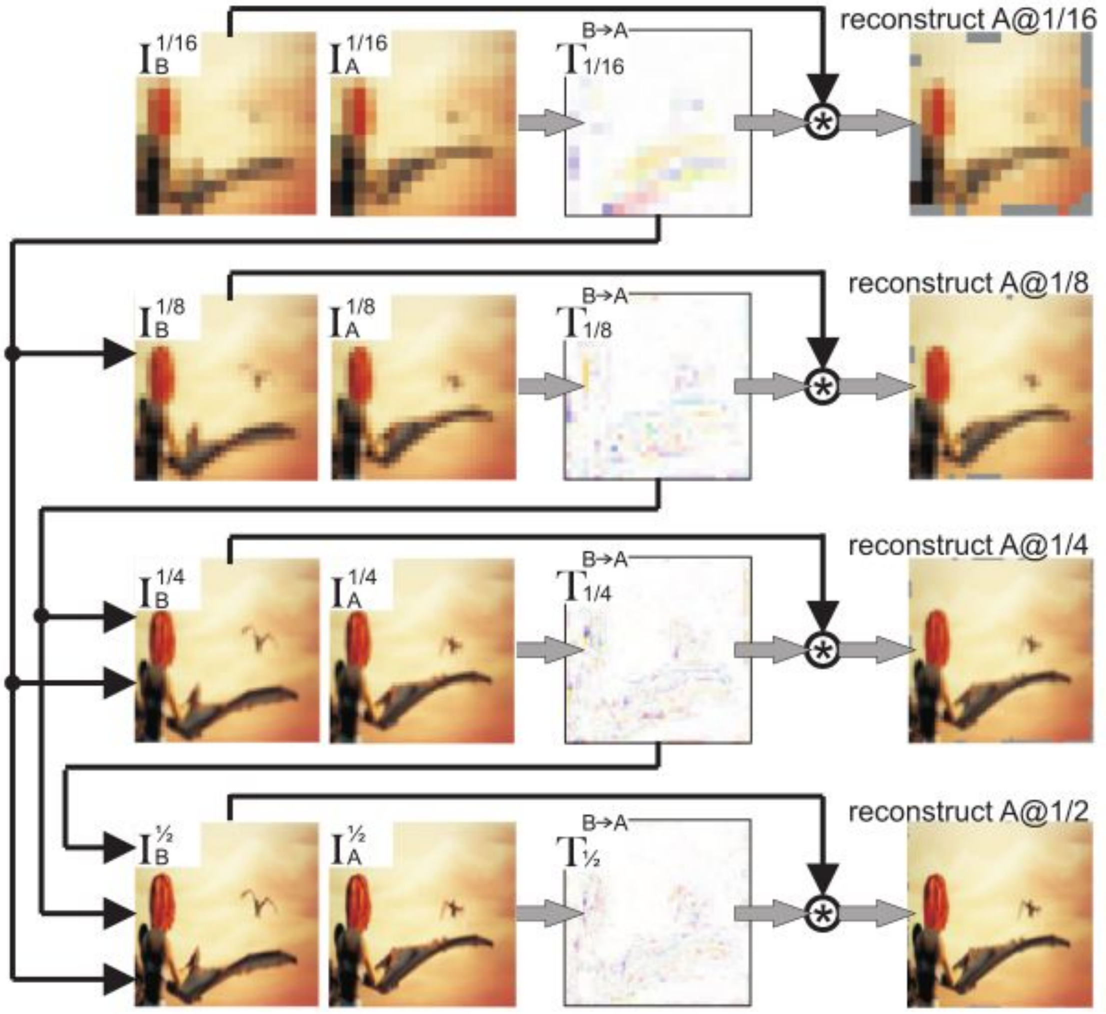
Multigrid PFF

voting for offset



$$\begin{bmatrix} v_x(i,j)\\ v_y(i,j)\end{bmatrix} = \sum_{x,y}\hat{T}_{ij,xy}\begin{bmatrix}x-i\\ y-j\end{bmatrix}$$

$$\begin{pmatrix}1.1\\ 1.4\end{pmatrix} = \begin{pmatrix}1\\ 1\end{pmatrix}*0.6+\begin{pmatrix}1\\ 2\end{pmatrix}*0.3+\begin{pmatrix}2\\ 2\end{pmatrix}*0.1$$



Multigrid Predictive Filter Flow (mgPFF)

Rather than 256x256x6400, with PFF of 11*11 kernel size for all scales, we have output with mgPFF as 256x256x121+128x128x121+64x64x121+32x32x121.

Multigrid Predictive Filter Flow (mgPFF)

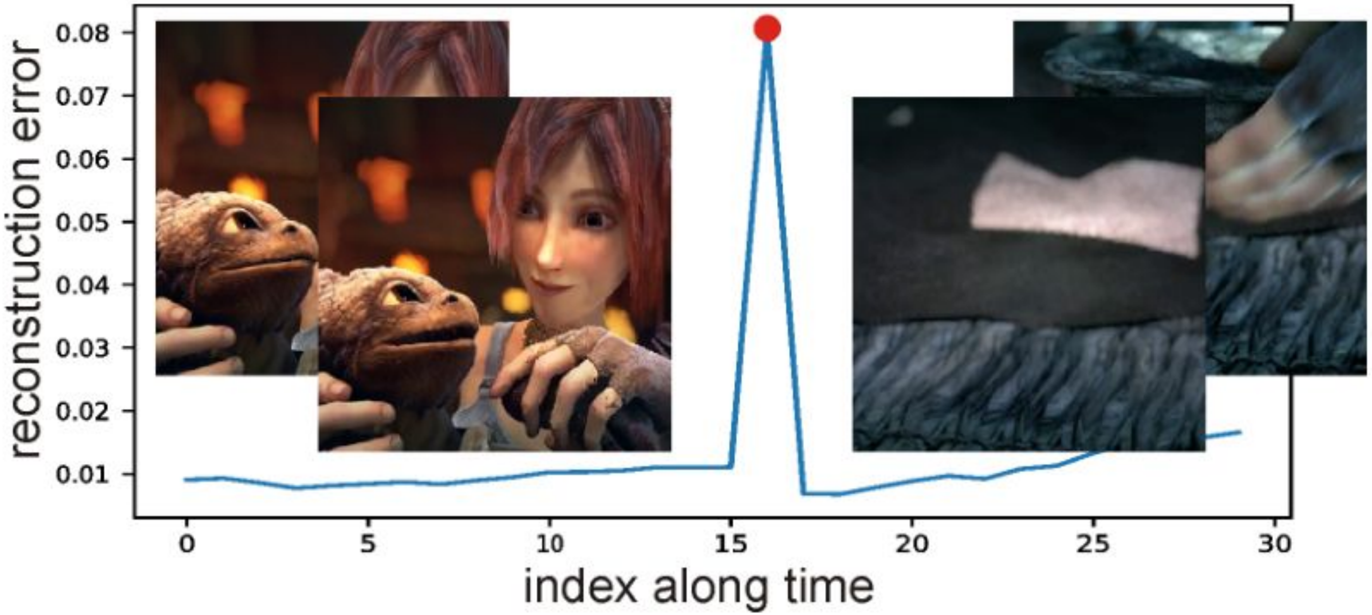
Rather than 256x256x6400, with PFF of 11*11 kernel size for all scales, we have output with mgPFF as 256x256x121+128x128x121+64x64x121+32x32x121.

With self-similarity across scales, sharing the weights to make it compact, resulting into a model of **4.6MB**.

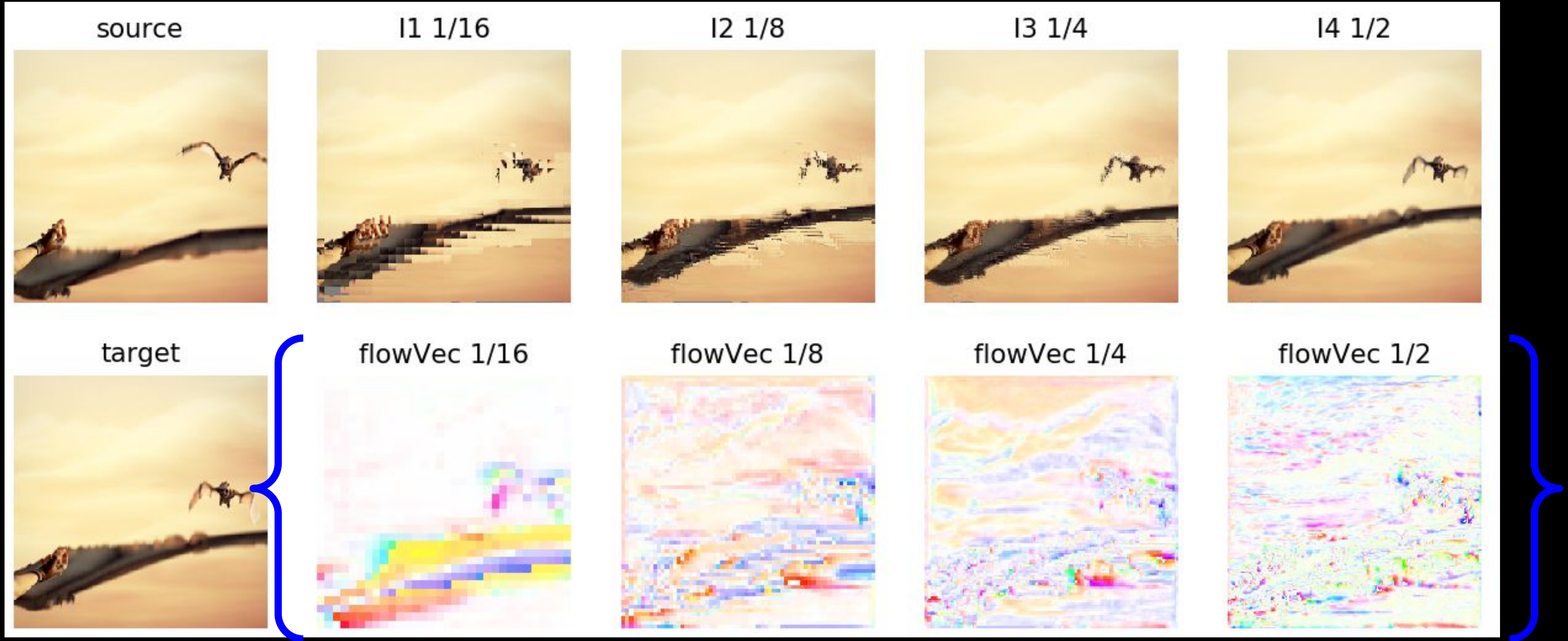
Multigrid Predictive Filter Flow (mgPFF)

training on free-form videos (e.g., the complete Sintel Movie).

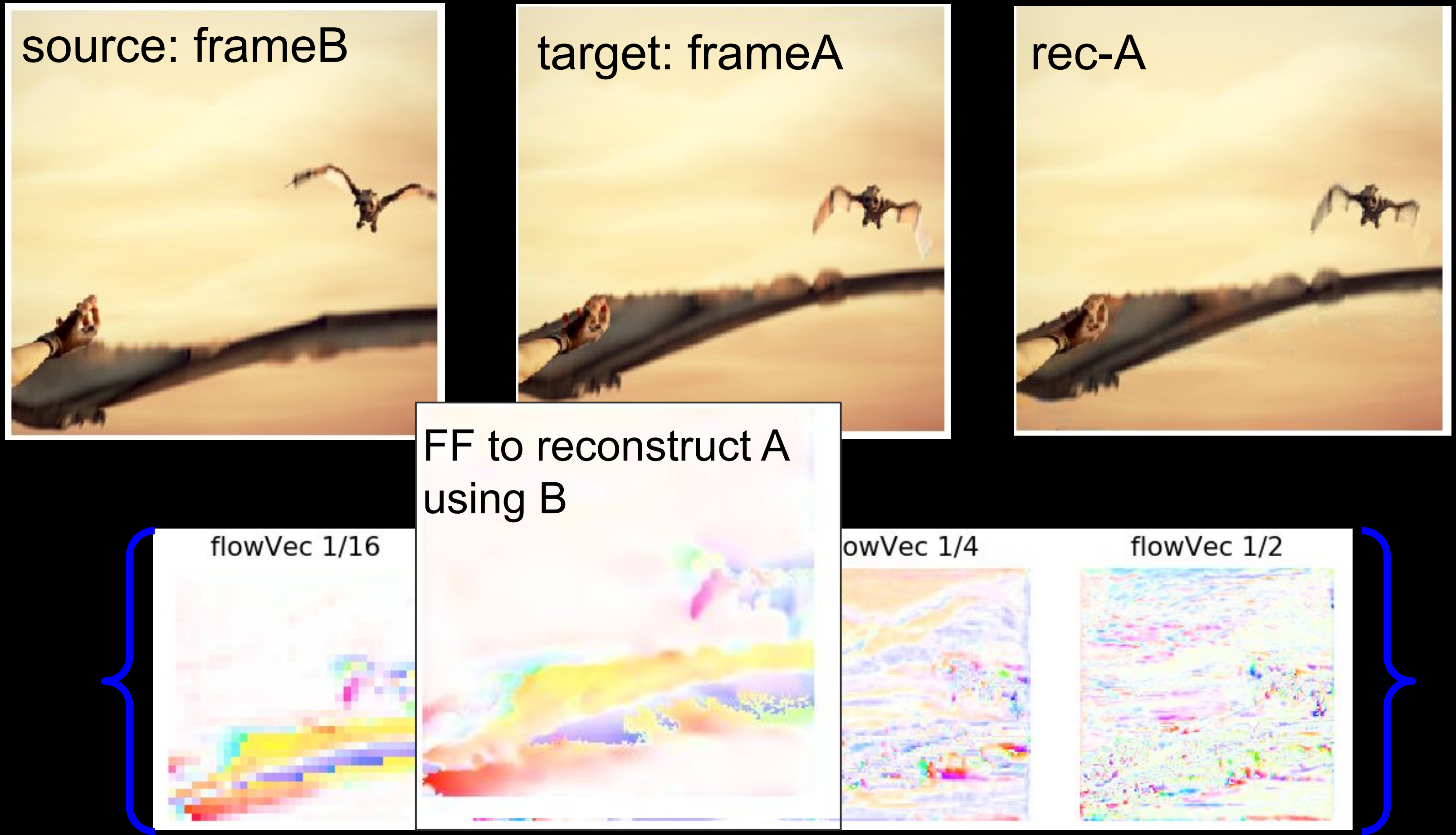
byproduct: video transition/shot detection



Multigrid Predictive Filter Flow (mgPFF)



Multigrid Predictive Filter Flow (mgPFF)



Applications of mgPFF

various tasks, for example--

1. transition/shot detection
2. video instance tracking, human pose tracking
3. long-range flow

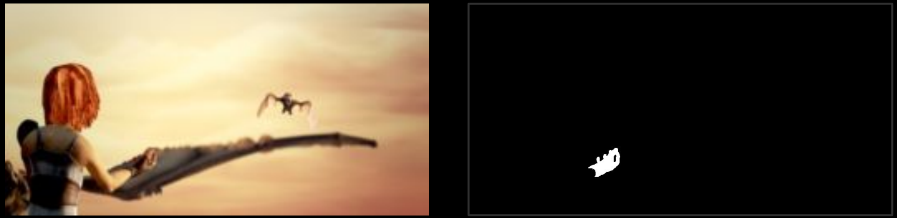
mgPFF for Instance Tracking

simply propagating the mask using the estimated flow

mgPFF for Instance Tracking

simply propagating the mask using the estimated flow

tracking ***right hand***



mgPFF for Instance Tracking

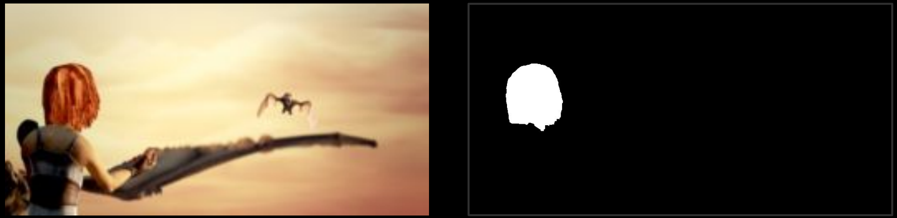
simply propagating the mask using the estimated flow

tracking ***bird***

mgPFF for Instance Tracking

simply propagating the mask using the estimated flow

tracking ***head***



mgPFF for Instance Tracking

simply propagating the mask using the estimated flow

benchmarking on the DAVIS dataset

mgPFF for Instance Tracking

K=3 [1, t-2, t-1] using first and previous two frames for tracking

mgPFF for Instance Tracking

K=1 [t-1] using the previous frame for tracking

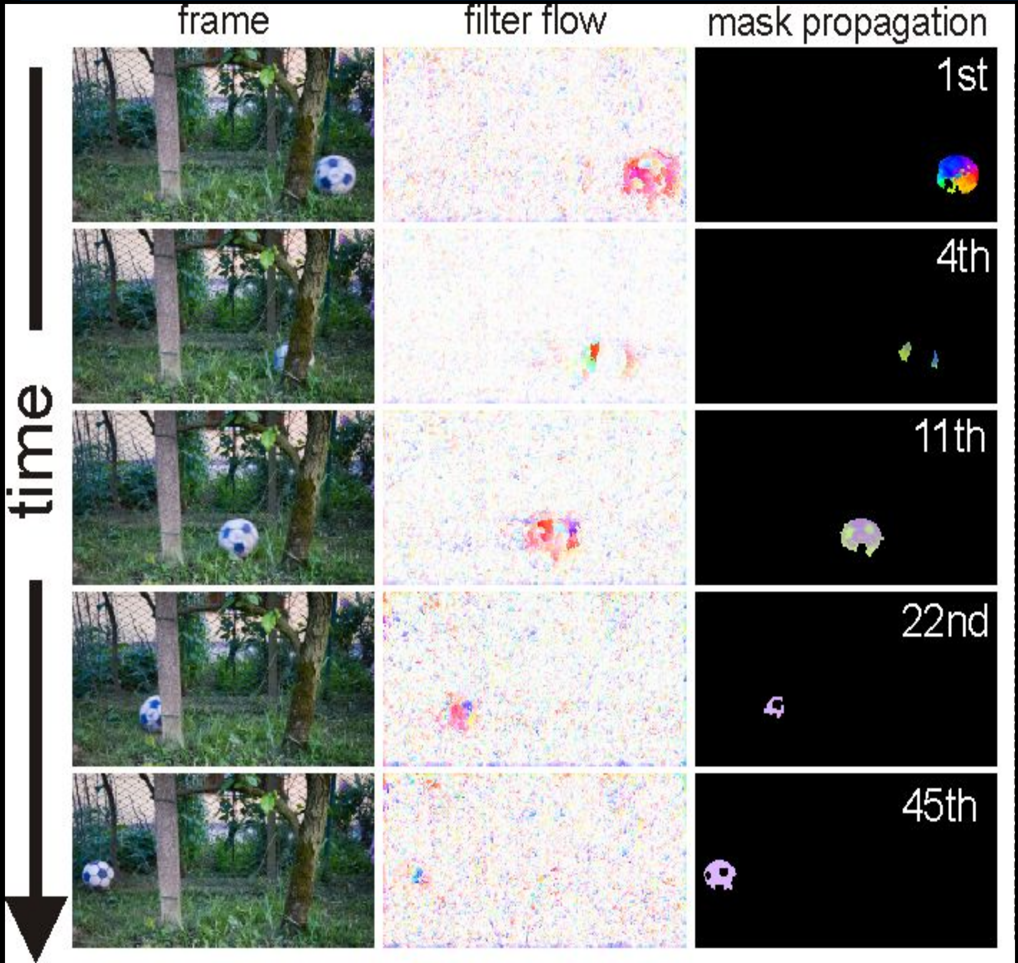
mgPFF for Instance Tracking

K=1 [1] using the first frame for tracking

mgPFF for Instance Tracking

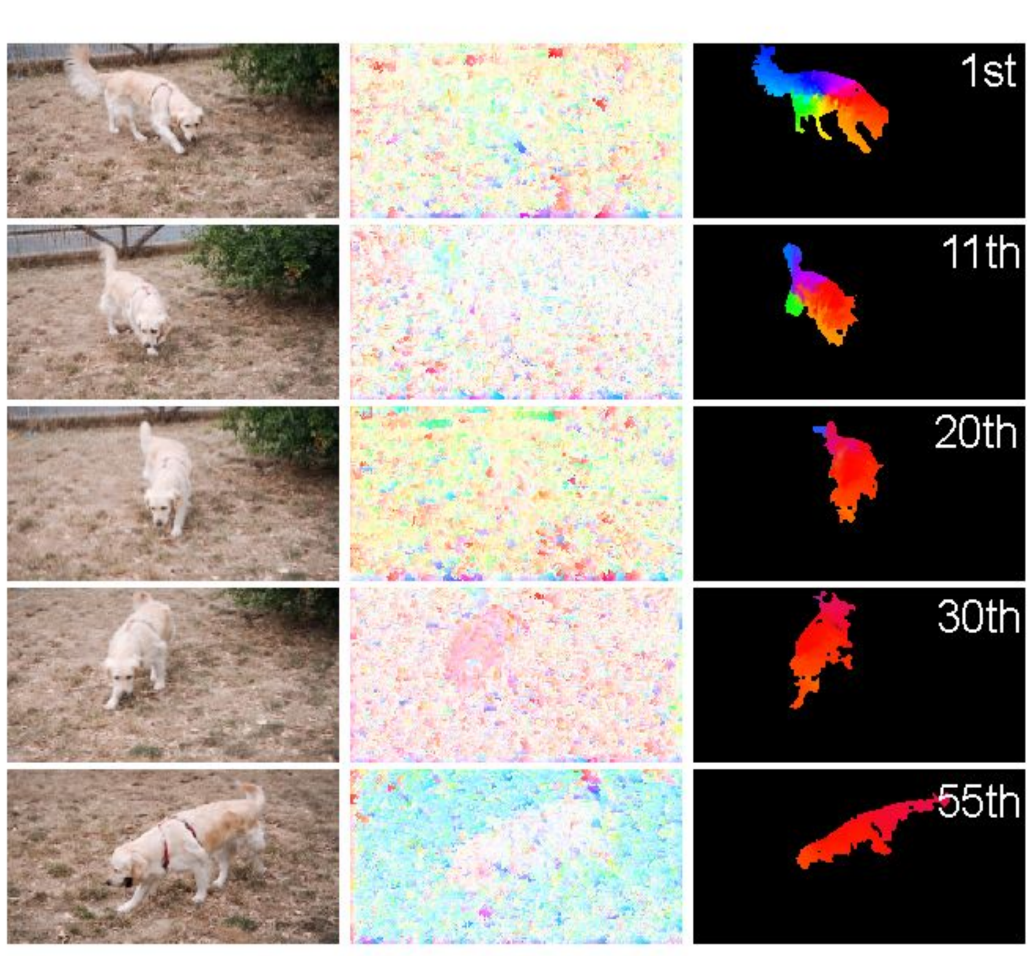
Method	Supervision	$\mathcal{J}$ (segments)		$\mathcal{F}$ (boundaries)	
		mean↑	recall↑	mean↑	recall↑
Identity	None	22.1	15.9	23.6	11.7
SIFTflow [46]	None	13.0	7.9	15.1	5.5
SIFTflow^{1st} [46]	None	33.0	–	35.0	–
FlowNet2 [29]	Synthetic	16.7	9.5	19.7	7.6
FlowNet2^{1st} [29]	Synthetic	26.7	–	25.2	–
DeepCluster^{1st} [9]	Self (1.3×10^6)	37.5	–	33.2	–
ColorPointer [91]	Self (9.0×10^7)	34.6	34.1	32.7	26.8
CycleTime^{1st} [94]	Self (3.7×10^7)	40.1	–	38.3	–
mgPFF (1st only)	Self (6.0×10^4)	31.6	29.5	36.2	30.8
mgPFF (K=1)		38.9	38.5	41.1	38.6
mgPFF^{1st} (K=1)		41.9	41.4	45.2	43.9
mgPFF^{1st} (K=3)		**42.2**	**41.8**	**46.9**	**44.4**

mgPFF for Instance Tracking



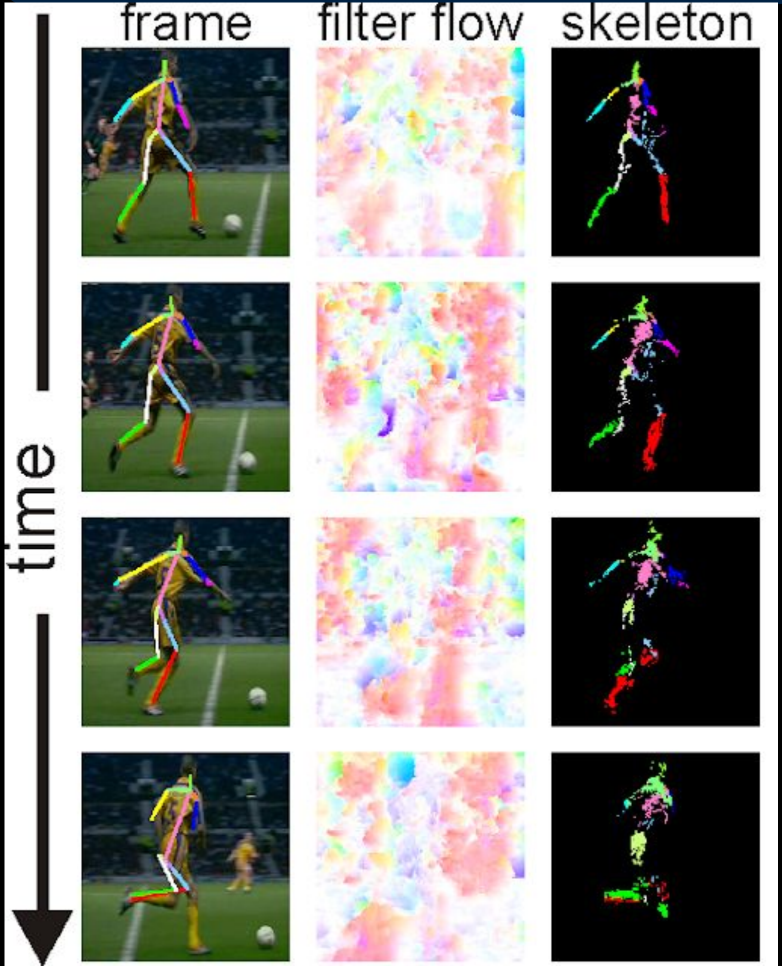
how it deals with heavy occlusion

mgPFF for Instance Tracking

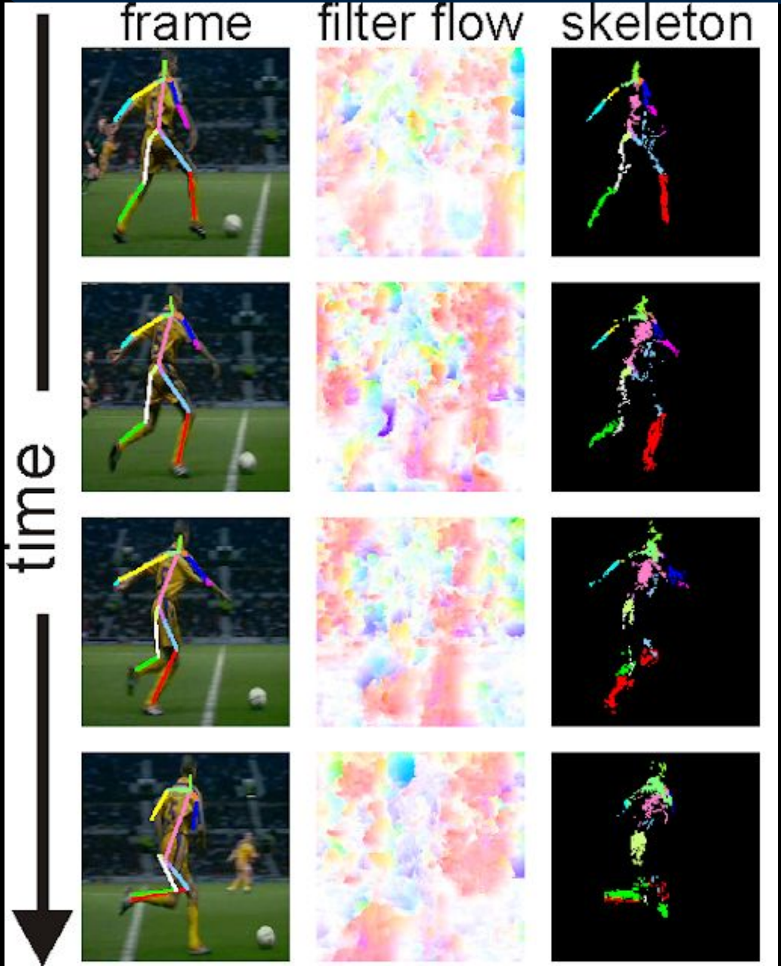


how it deals with large deformation

mgPFF for Pose Tracking



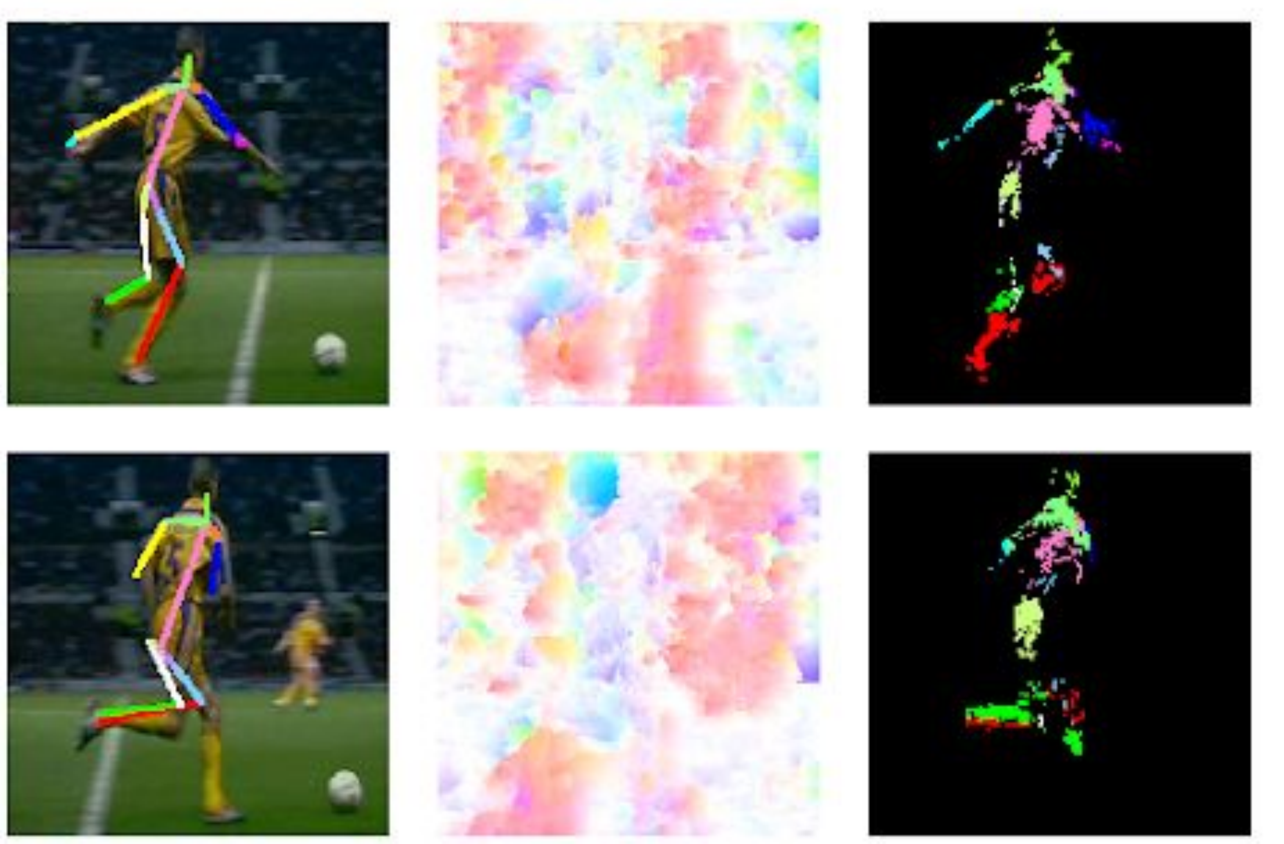
mgPFF for Pose Tracking



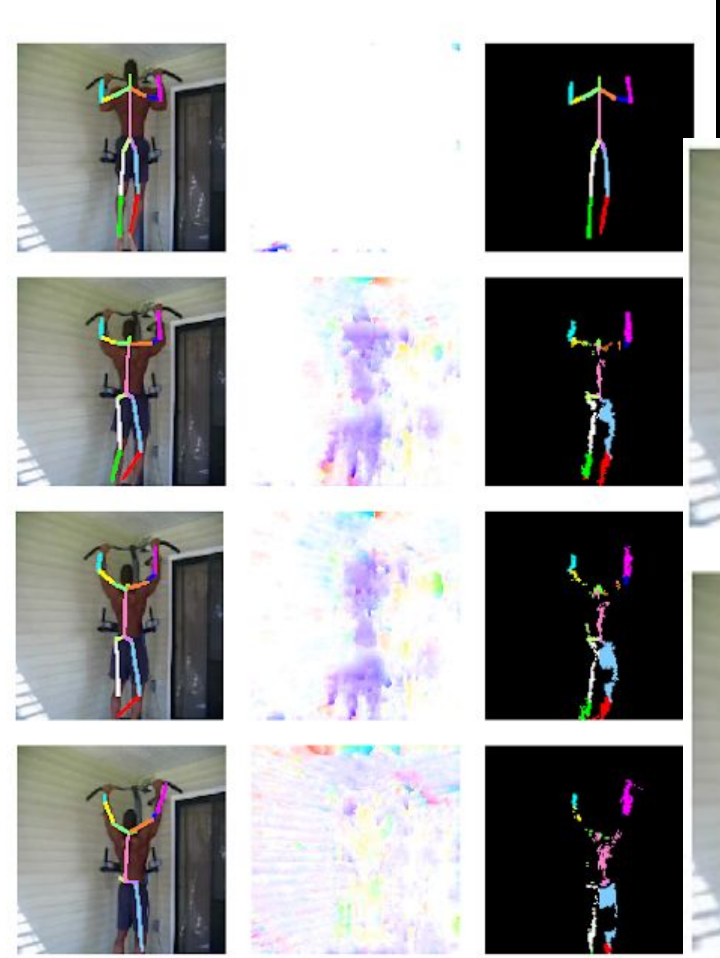
Method / PCK↑	@0.1	@0.2
fully-supervised [84]	68.7	92.1
Identity	43.1	64.5
SIFTflow^{1st} [46]	49.0	68.6
FlowNet2 [29]	45.2	62.9
DeepCluster^{1st} [9]	43.2	66.9
ColorPointer [91]	45.2	69.6
CycleTime^{1st} [94]	57.3	78.1
mgPFF	49.3	72.8
mgPFF^{1st}	55.6	77.1
mgPFF+ft	52.7	75.1
mgPFF+ft^{1st}	**58.4**	**78.1**

mgPFF for Pose Tracking

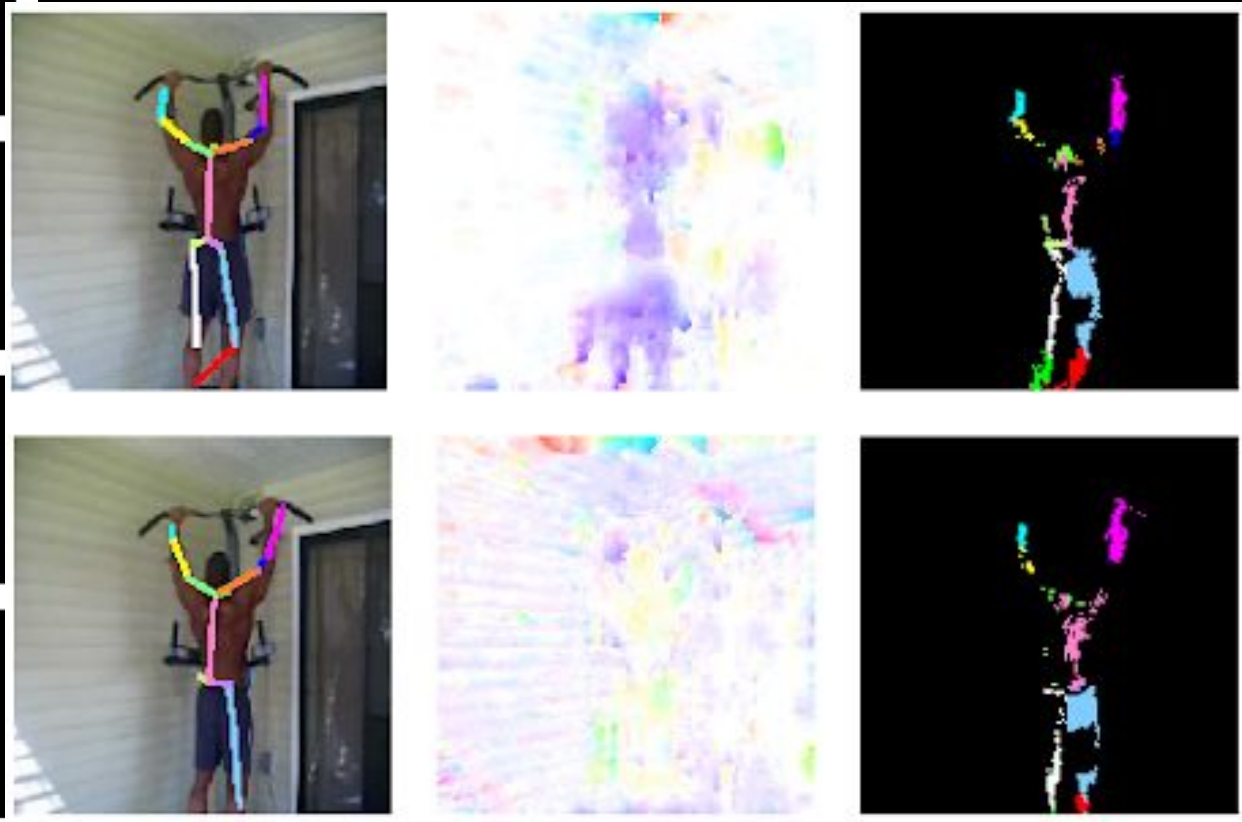
occlusion on the knees



mgPFF for Pose Tracking

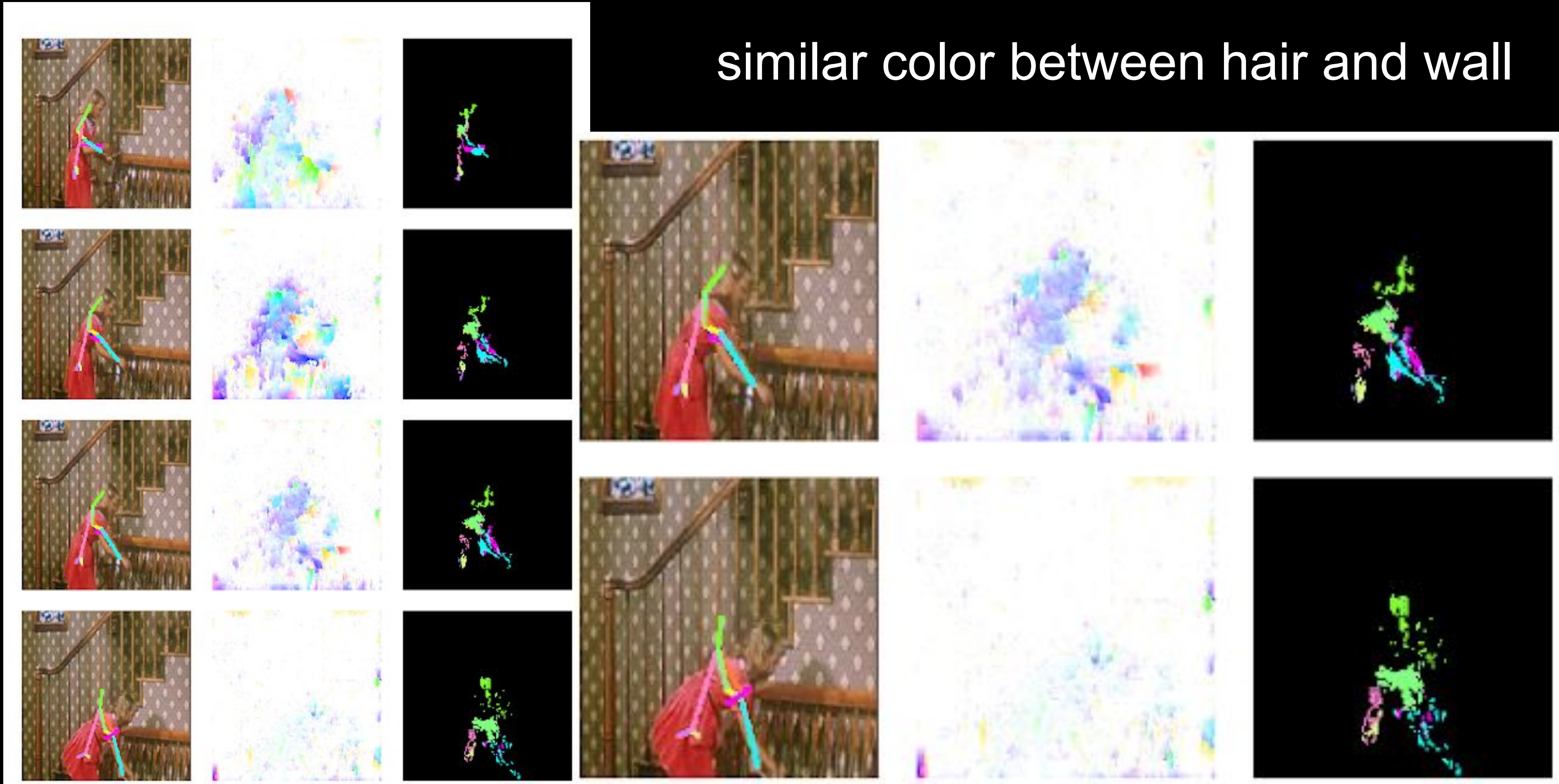


joints moving out of the box



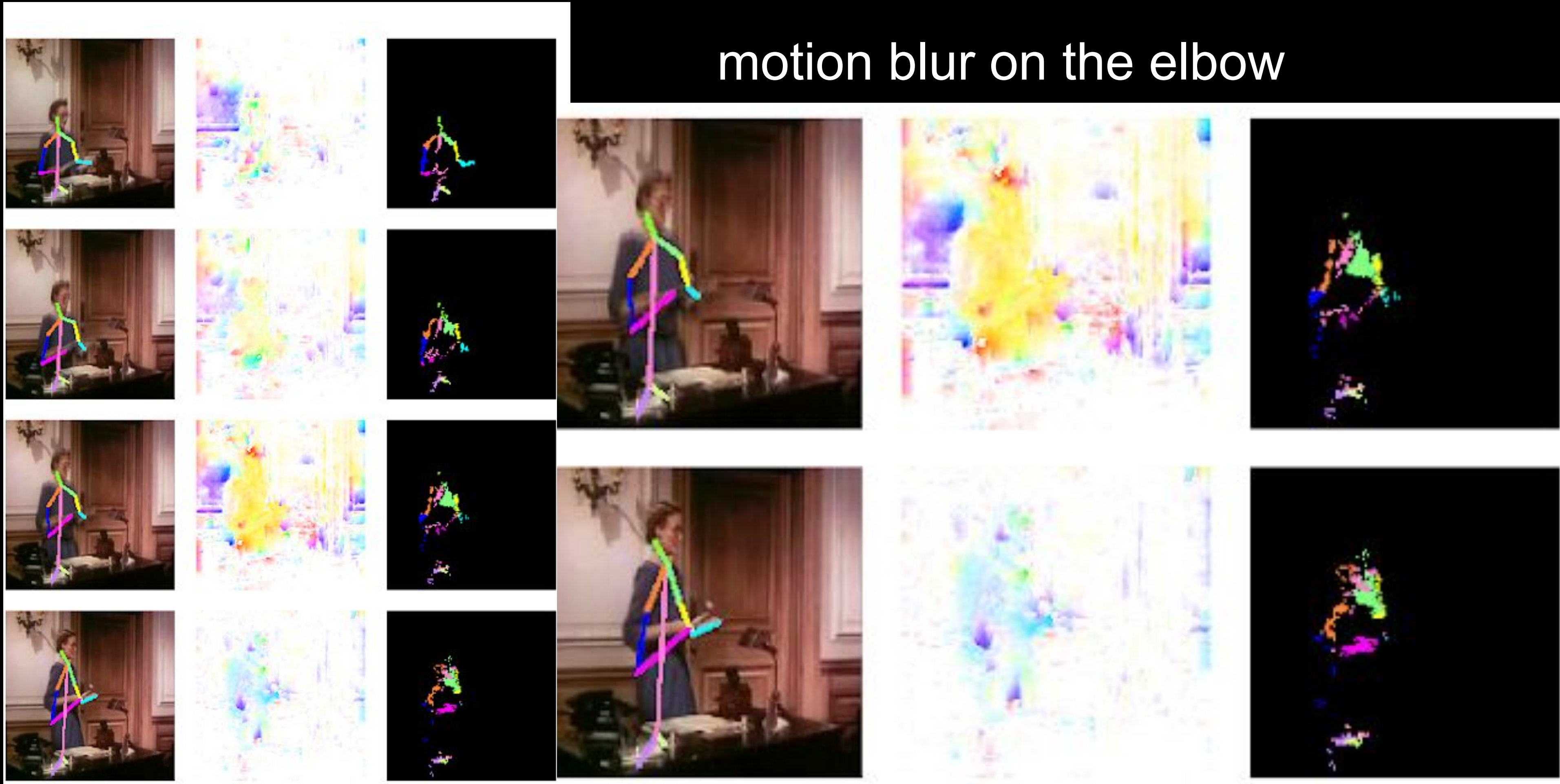
mgPFF for Pose Tracking

similar color between hair and wall

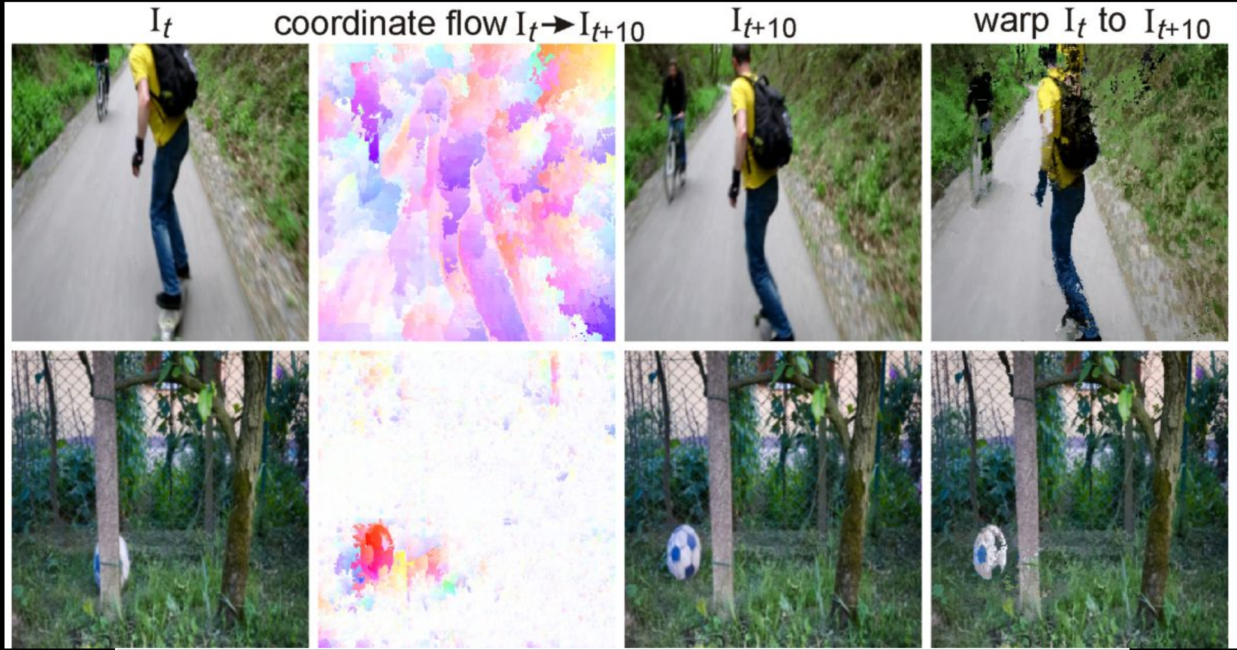


mgPFF for Pose Tracking

motion blur on the elbow



mgPFF for Long-Range Flow



method/error↓	5-Frame	10-Frame
Identity	82.0	97.7
Optical Flow (FlowNet2) [29]	62.4	90.3
CycleTime [94]	60.4	76.4
mgPFF	**7.32**	**8.83**

Summary: mgPFF for Video Mining

1. unsupervised learning framework on free-form videos;
2. compact model (4.6MB), easy training, fast computation;
3. better perf. of video tracking, great power for long-range flow;
4. interpretable in terms of decision making (per-pixel tracking);

Summary: mgPFF for Video Mining

1. unsupervised learning framework on free-form videos;
2. compact model (4.6MB), easy training, fast computation;
3. better perf. of video tracking, great power for long-range flow;
4. interpretable in terms of decision making (per-pixel tracking);
5. reminiscent of a variety of flow-based tasks

   video compression, frame interpolation, activity/action cls., optical flow, etc.

Summary: mgPFF for Video Mining

1. unsupervised learning framework on free-form videos;
2. compact model (4.6MB), easy training, fast computation;
3. better perf. of video tracking, great power for long-range flow;
4. interpretable in terms of decision making (per-pixel tracking);
5. reminiscent of a variety of flow-based tasks
   video compression, frame interpolation, activity/action cls., optical flow, etc.
6. interpretable model for good (transparent decision making)
   e.g., medical image enhancement

PFF for Single Image Reconstruction

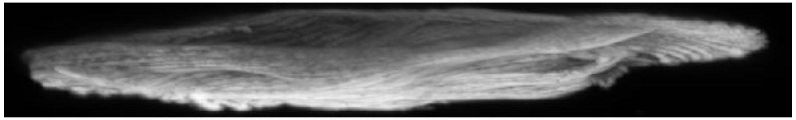
$$\mathbf{I}_2 \approx \hat{\mathbf{T}}\mathbf{I}_1, \begin{cases} \boxed{\hat{\mathbf{T}} \equiv f_{\mathbf{w}}(\mathbf{I}_1)} \\ \hat{\mathbf{T}} \equiv f_{\mathbf{w}}(\mathbf{I}_1, \mathbf{I}_2) \end{cases}$$

PFF for Single Image Reconstruction

$$\mathbf{I}_2 \approx \hat{\mathbf{T}}\mathbf{I}_1, \begin{cases} \boxed{\hat{\mathbf{T}} \equiv f_{\mathbf{w}}(\mathbf{I}_1)} \\ \hat{\mathbf{T}} \equiv f_{\mathbf{w}}(\mathbf{I}_1, \mathbf{I}_2) \end{cases}$$

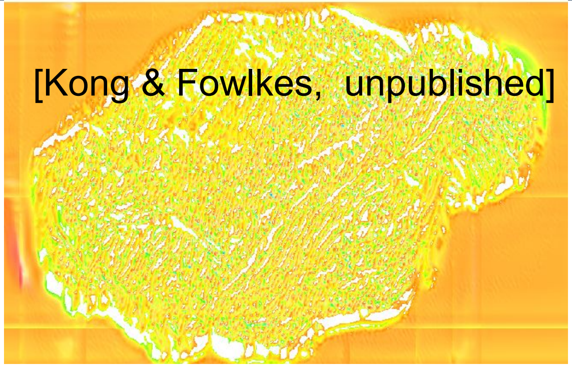
[Kong & Fowlkes, unpublished]

original size-view image

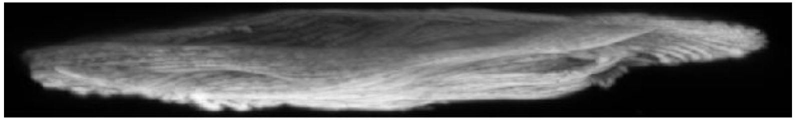


PFF for Single Image Reconstruction

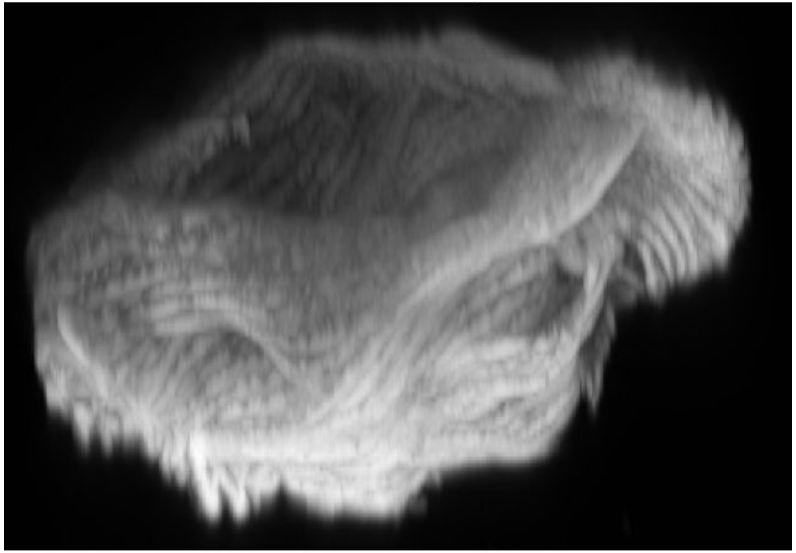
$$\mathbf{I}_2 \approx \hat{\mathbf{T}}\mathbf{I}_1, \begin{cases} \boxed{\hat{\mathbf{T}} \equiv f_{\mathbf{w}}(\mathbf{I}_1)} \\ \hat{\mathbf{T}} \equiv f_{\mathbf{w}}(\mathbf{I}_1, \mathbf{I}_2) \end{cases}$$



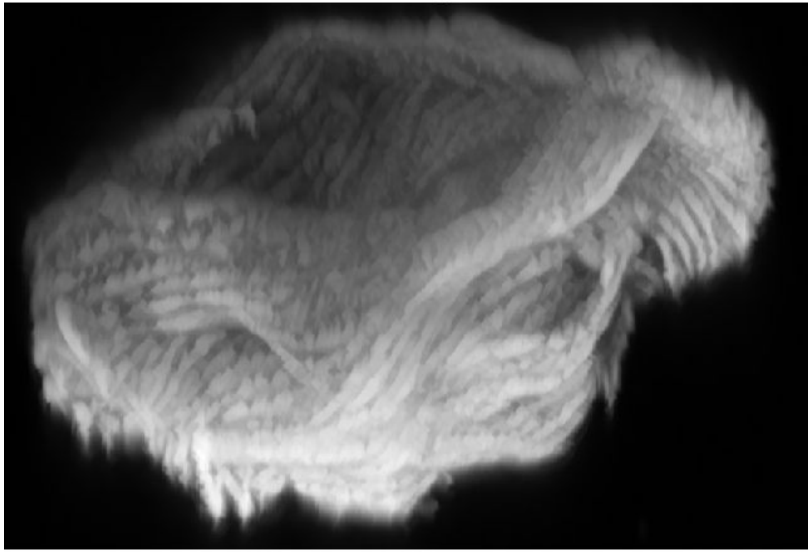
original size-view image



stretched side-view image



enhanced image



PFF for Single Image Reconstruction

$$\mathbf{I}_2 \approx \hat{\mathbf{T}}\mathbf{I}_1, \begin{cases} \boxed{\hat{\mathbf{T}} \equiv f_{\mathbf{w}}(\mathbf{I}_1)} \\ \hat{\mathbf{T}} \equiv f_{\mathbf{w}}(\mathbf{I}_1, \mathbf{I}_2) \end{cases}$$

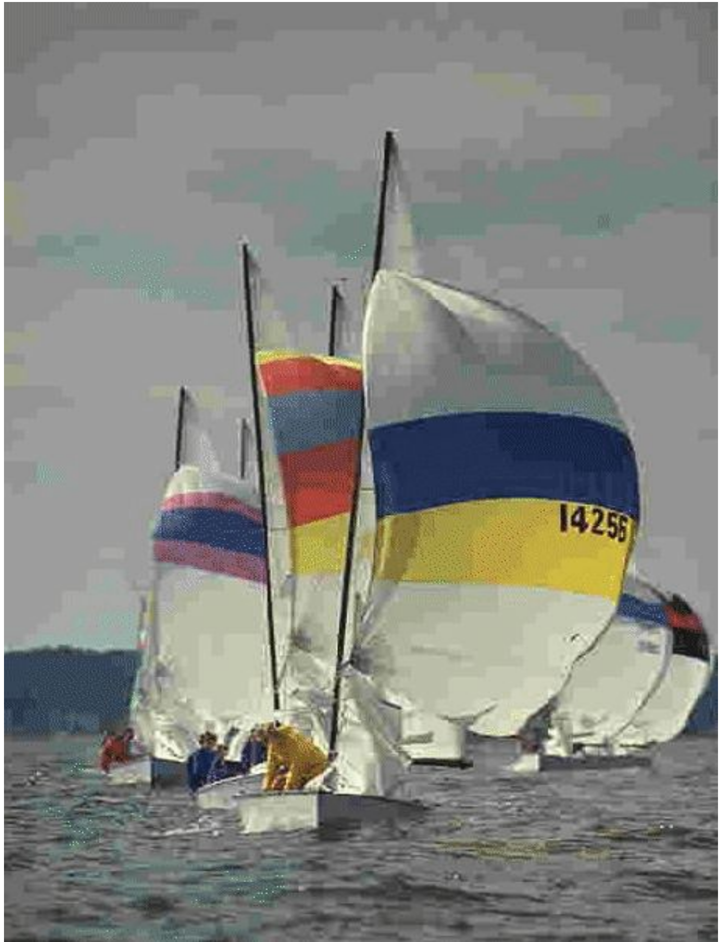
non-uniform deblur

[Kong & Fowlkes, 2018]

PFF for Single Image Reconstruction

$$\mathbf{I}_2 \approx \hat{\mathbf{T}}\mathbf{I}_1, \begin{cases} \boxed{\hat{\mathbf{T}} \equiv f_{\mathbf{w}}(\mathbf{I}_1)} \\ \hat{\mathbf{T}} \equiv f_{\mathbf{w}}(\mathbf{I}_1, \mathbf{I}_2) \end{cases}$$

lossy compression artifact reduction



[Kong & Fowlkes, 2018]

PFF for Single Image Reconstruction

$$\mathbf{I}_2 \approx \hat{\mathbf{T}}\mathbf{I}_1, \begin{cases} \boxed{\hat{\mathbf{T}} \equiv f_{\mathbf{w}}(\mathbf{I}_1)} \\ \hat{\mathbf{T}} \equiv f_{\mathbf{w}}(\mathbf{I}_1, \mathbf{I}_2) \end{cases}$$

single image super-resolution

[Kong & Fowlkes, 2018]

Summary: mgPFF for Video Mining

1. unsupervised learning framework on free-form videos;
2. compact model (4.6MB), easy training, fast computation;
3. better perf. of video tracking, great power for long-range flow;
4. interpretable in terms of decision making (per-pixel tracking);
5. reminiscent of a variety of flow-based tasks

   video compression, frame interpolation, activity/action cls., optical flow, etc.
6. interpretable model for good (transparent decision making)

   e.g., medical image enhancement
7. abundant future work

   combining higher-level info., mobile dev., etc.

Outline of Video Mining

1. Unsupervised Learning with Multigrid Predictive Filter Flow

   video inst. seg./tracking, pose tracking, long-range flow, video shot det.

2. tba
3. tba
4. Conclusion with discussion

Conclusion

1. Learning with videos in a more affordable way (not much supervision required)
2. low-vision mining to mid-level application, high-level learning

With videos, a lot is happening

Conclusion

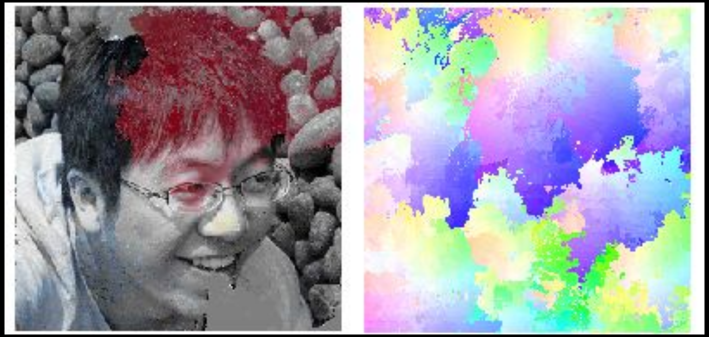
1. Learning with videos in a more affordable way (not much supervision required)
2. low-vision mining to mid-level application, high-level learning

With videos, a lot is happening; some future explorations --

- visual commonsense/knowledge
  - affordance, correspondence, parts, etc.
- better human-machine intersection (assistive robots)
- better intelligent systems

Thanks

Thanks



Shu Kong & Charless Fowlkes, 2019

Thanks

Q&A