Beyond Correlation Filters: Learning Continuous Convolution Operators for Visual Tracking

Martin Danelljan, Andreas Robinson,
Fahad Shahbaz Khan, Michael Felsberg

LINKÖPING UNIVERSITY

“tracking itself is by and large a solved problem”

[Jianbo Shi & Carlo Tomasi CVPR 1994]

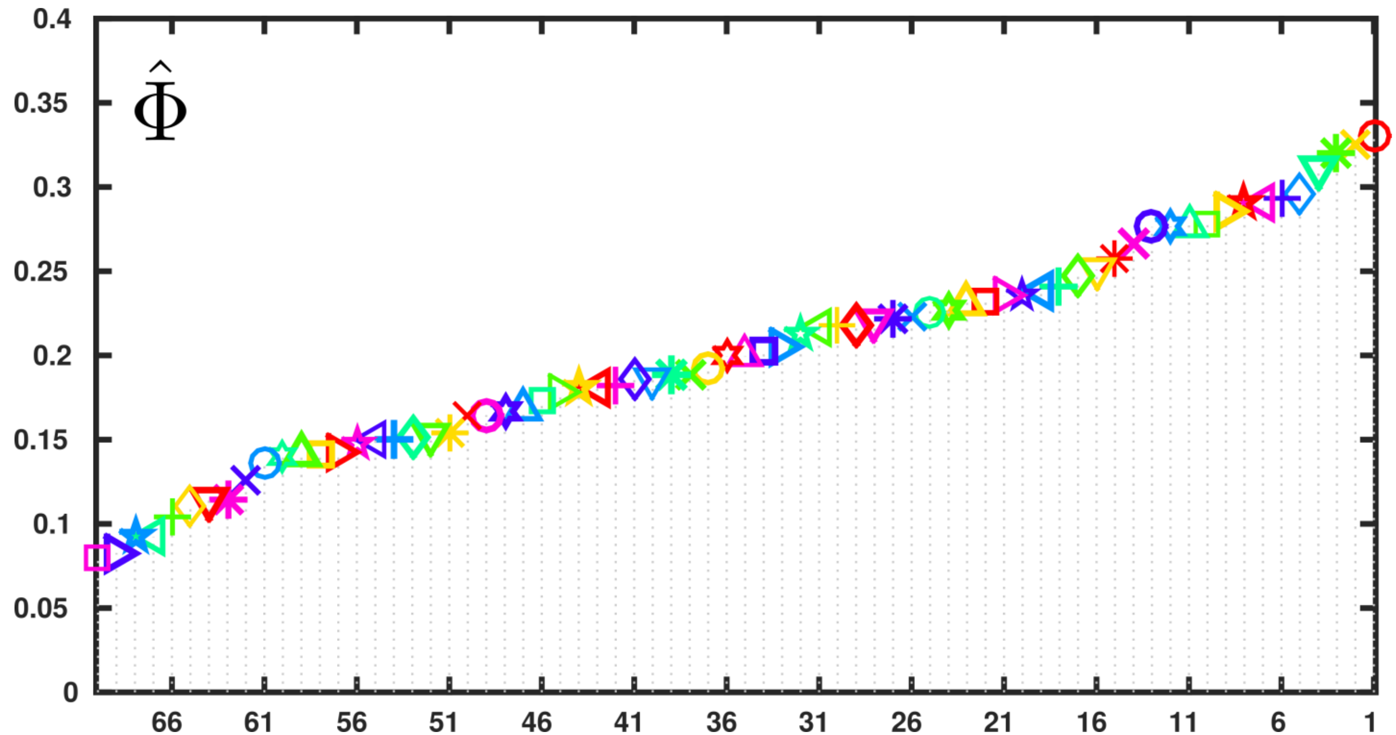
LINKÖPING UNIVERSITY

[Matej et al., ECCV VOT workshop 2016]

“tracking itself is by and large a solved problem”

[Jianbo Shi & Carlo Tomasi CVPR 1994]

Visual Object Tracking (VOT) 2016 challenge results:

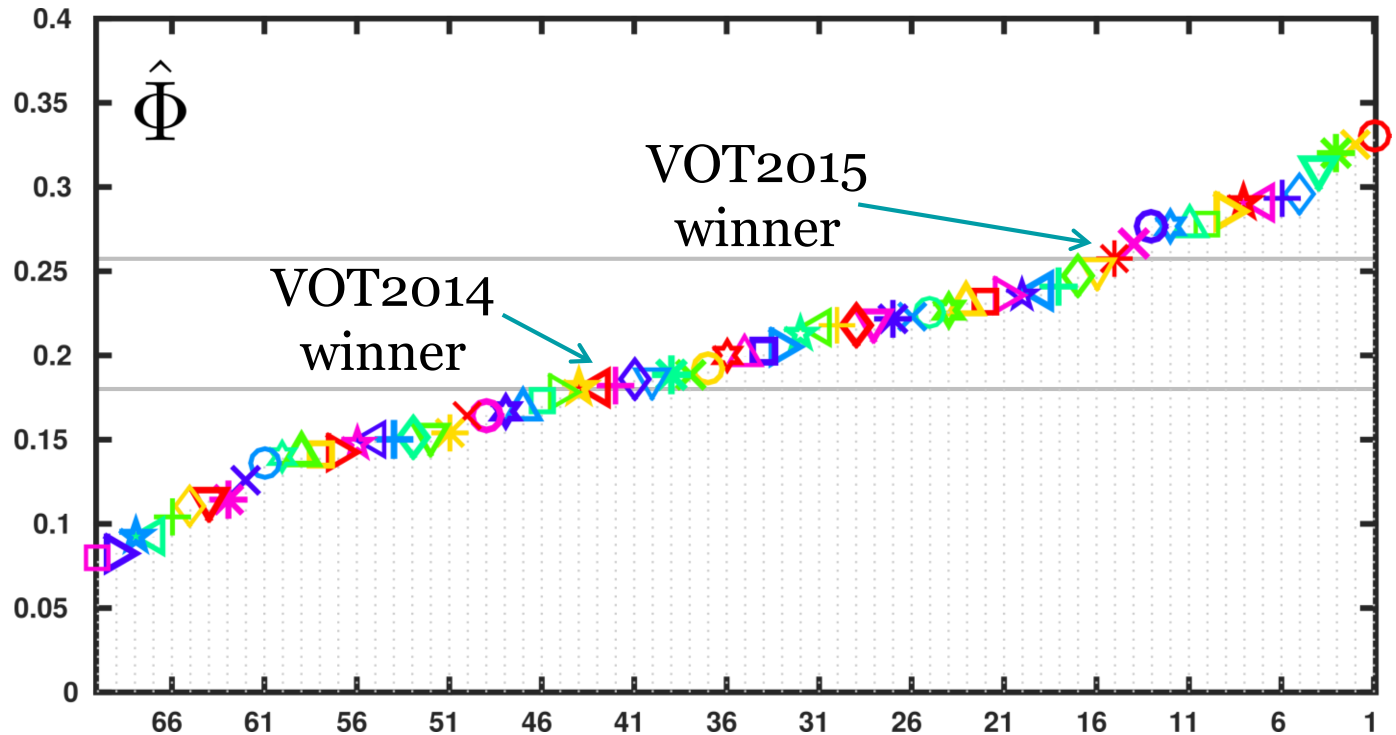


[Matej et al., ECCV VOT workshop 2016]

"tracking itself is by and large a solved problem"

[Jianbo Shi & Carlo Tomasi CVPR 1994]

Visual Object Tracking (VOT) 2016 challenge results:



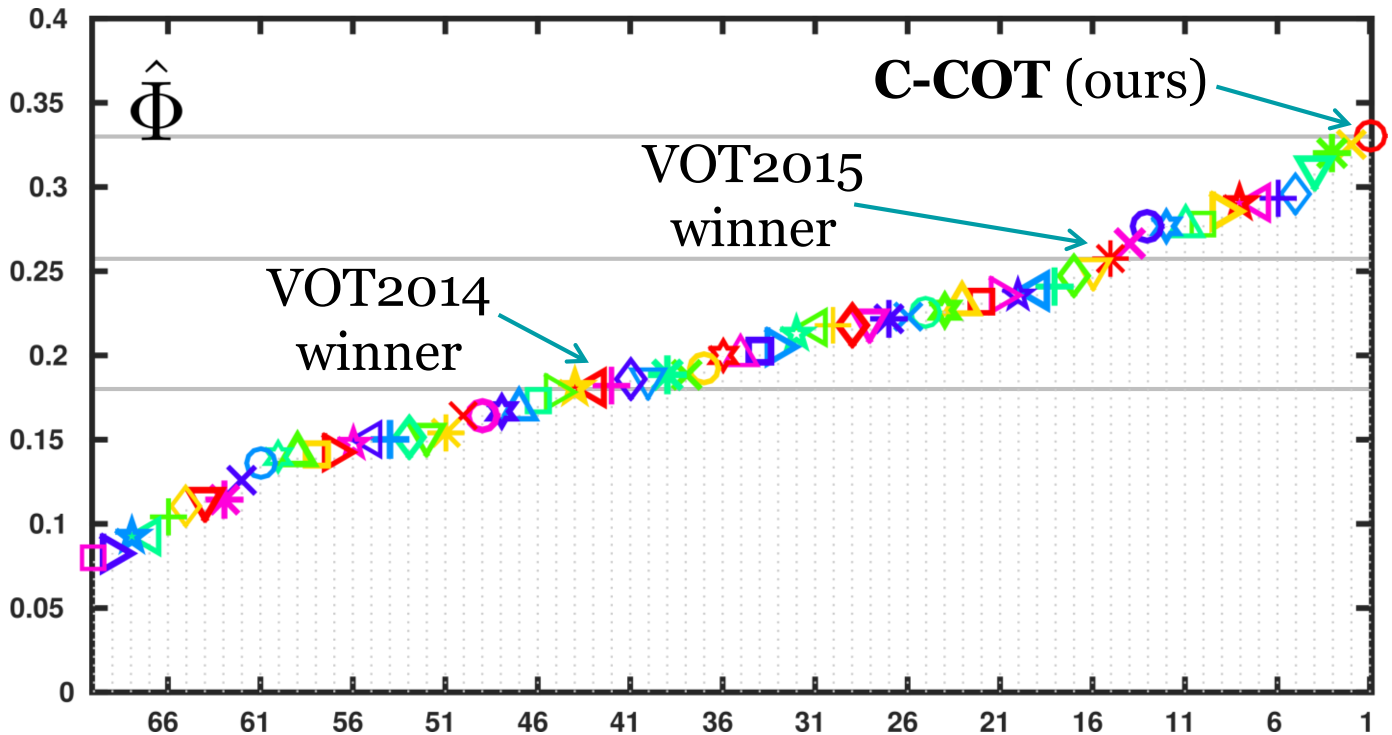
[Matej et al., ECCV VOT workshop 2016]

LINKÖPING UNIVERSITY

"tracking itself is by and large a solved problem"

[Jianbo Shi & Carlo Tomasi CVPR 1994]

Visual Object Tracking (VOT) 2016 challenge results:



[Matej et al., ECCV VOT workshop 2016]

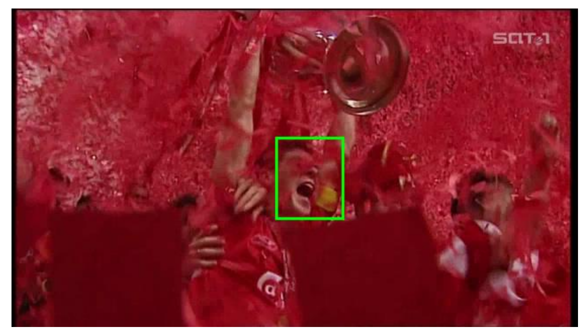
Tracking Challenges

Tracking Challenges

Blur

Appearance Change

Occlusion

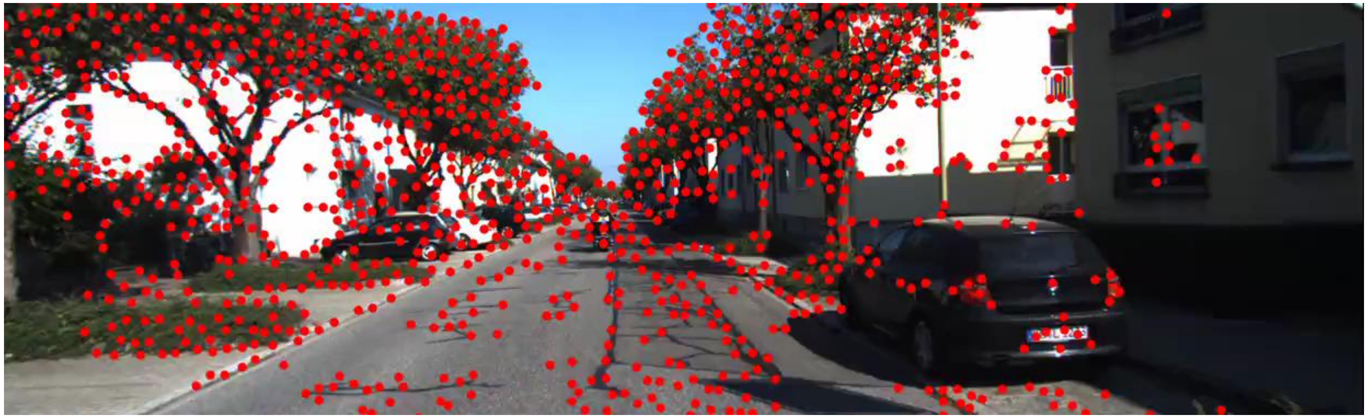


Clutter

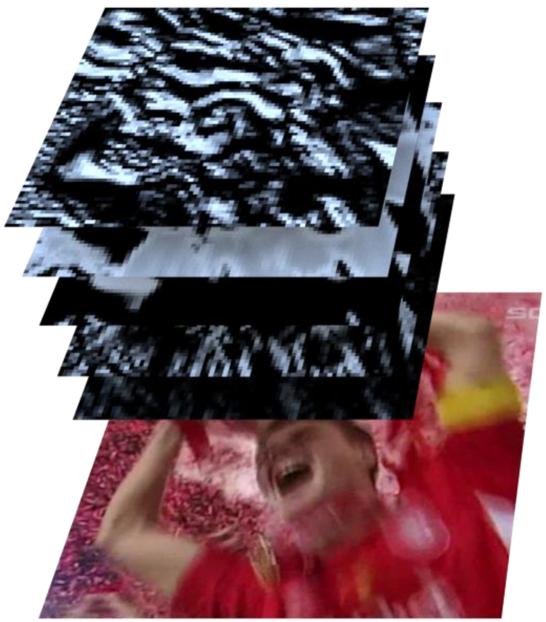
Feature Point Tracking

Our C-COT (discriminative)

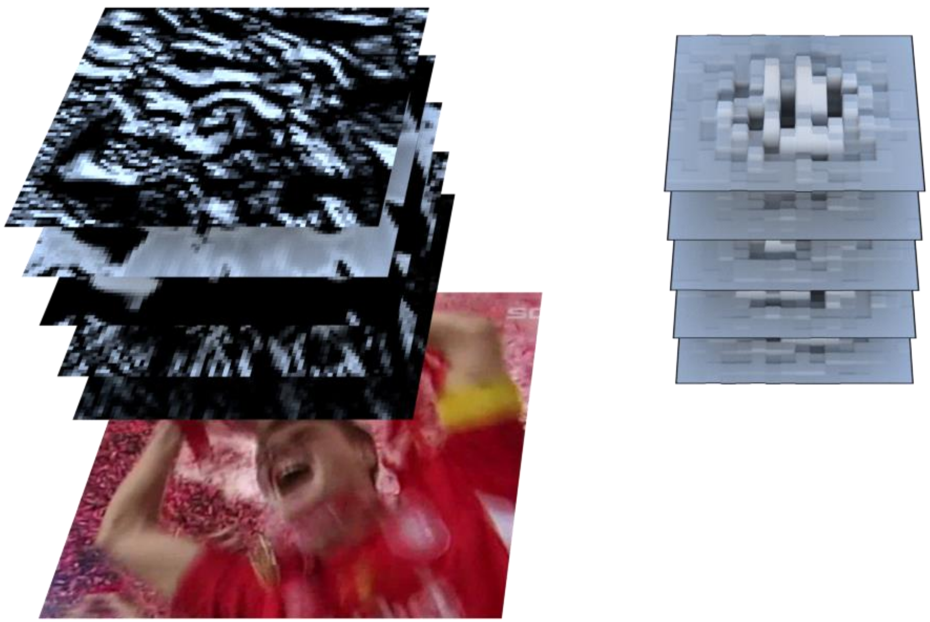
KLT (generative)



Discriminative Correlation Filters (DCF)



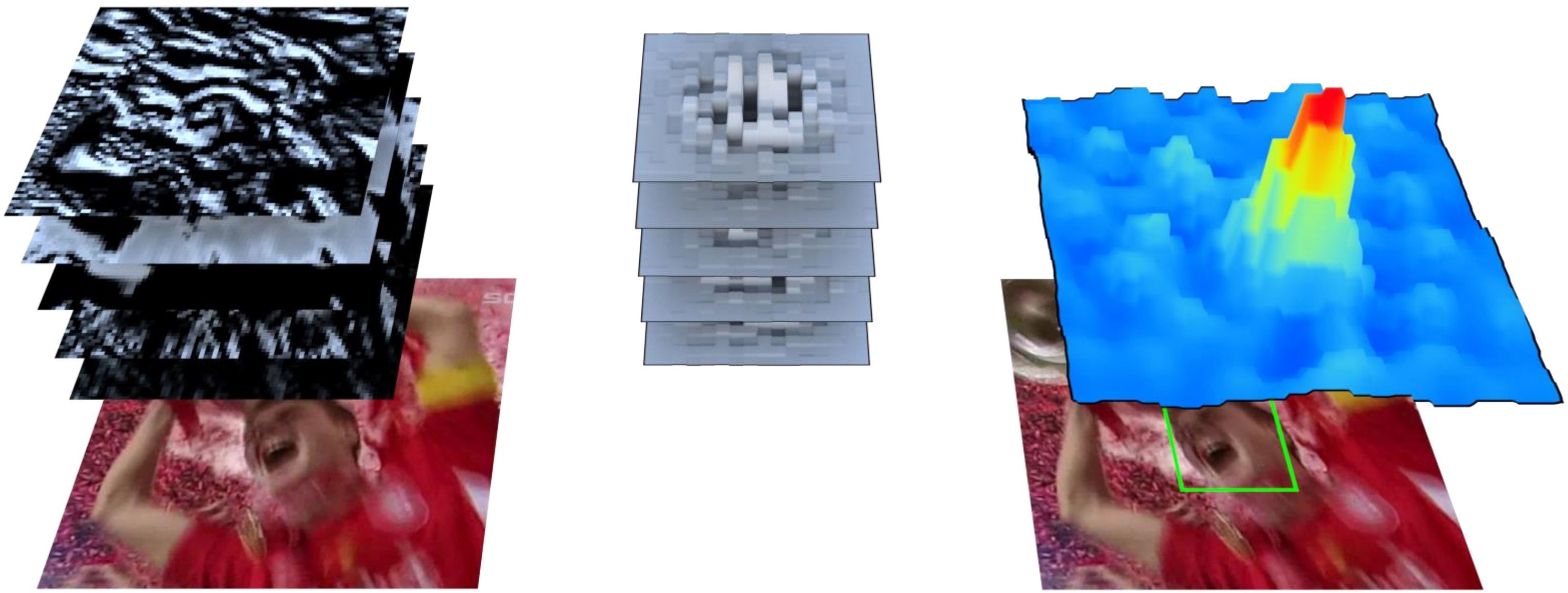
Discriminative Correlation Filters (DCF)



Discriminative Correlation Filters (DCF)

Discriminative Correlation Filters (DCF)

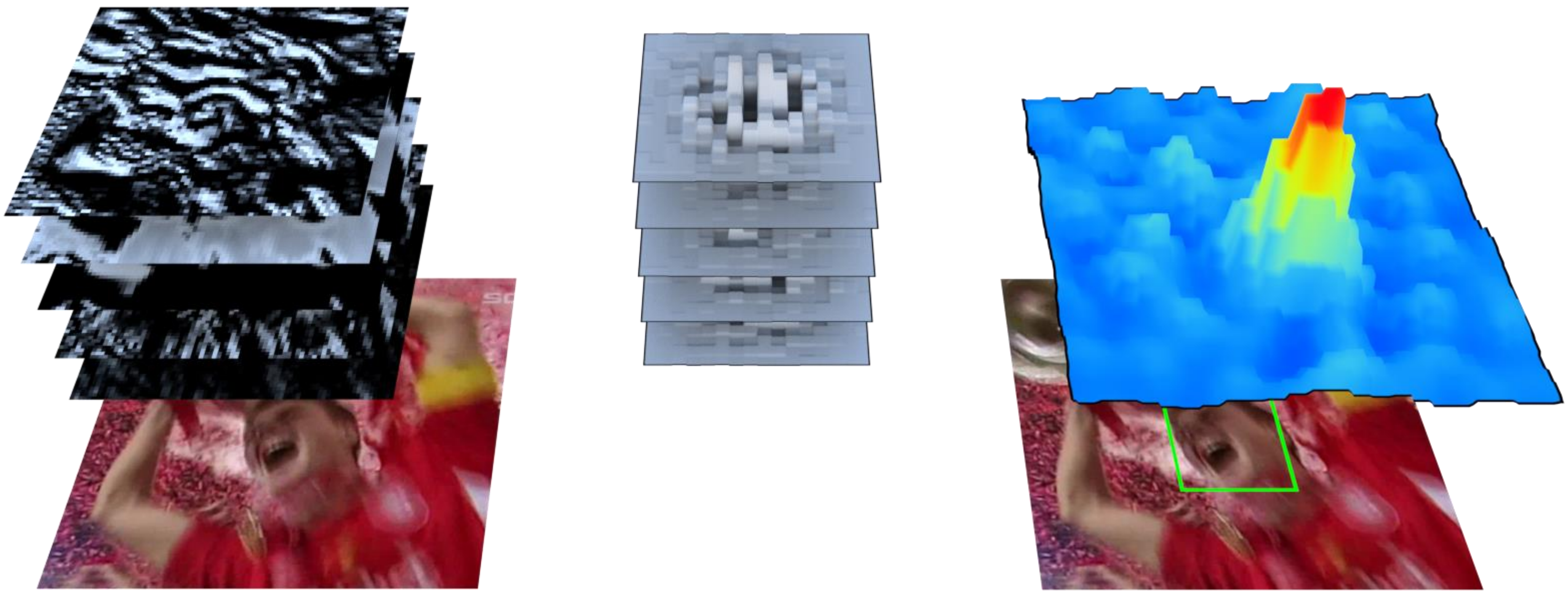
Limitations:



Discriminative Correlation Filters (DCF)

Limitations:

Single-resolution feature map

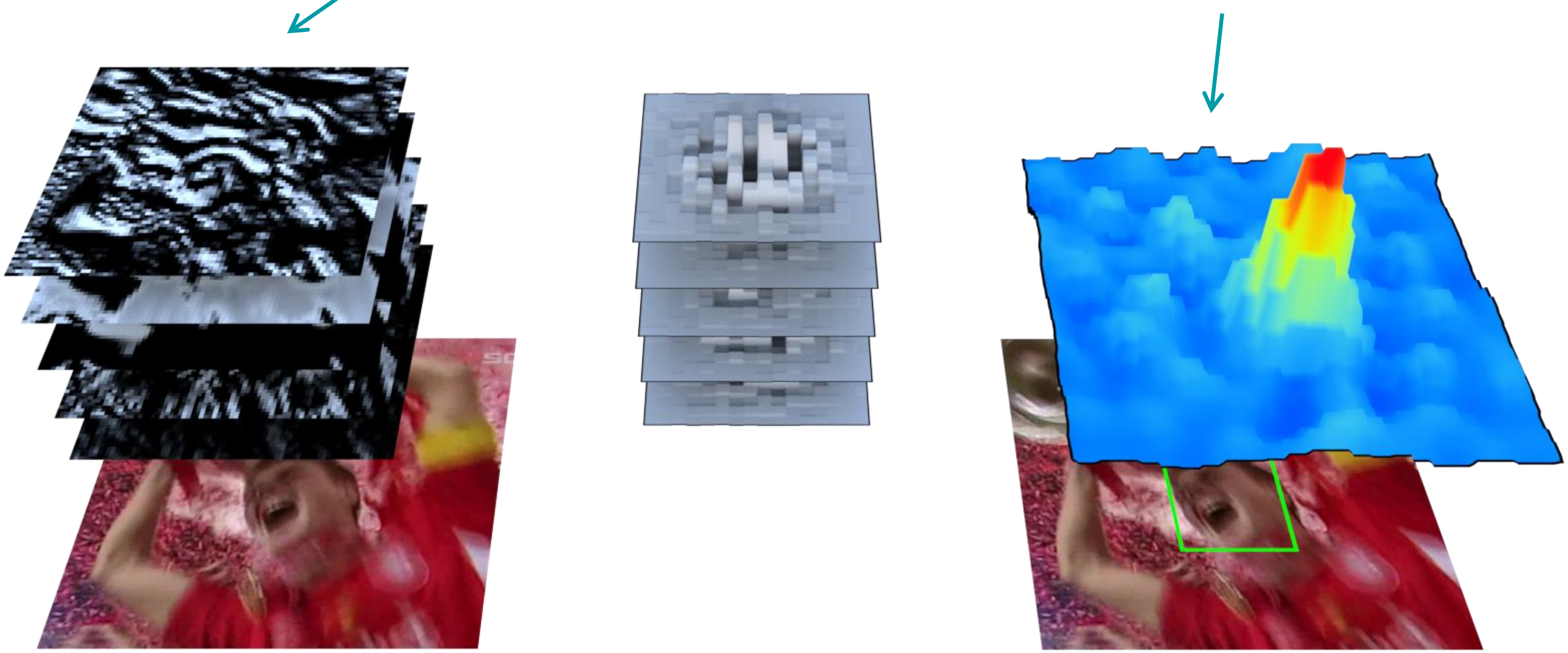


Discriminative Correlation Filters (DCF)

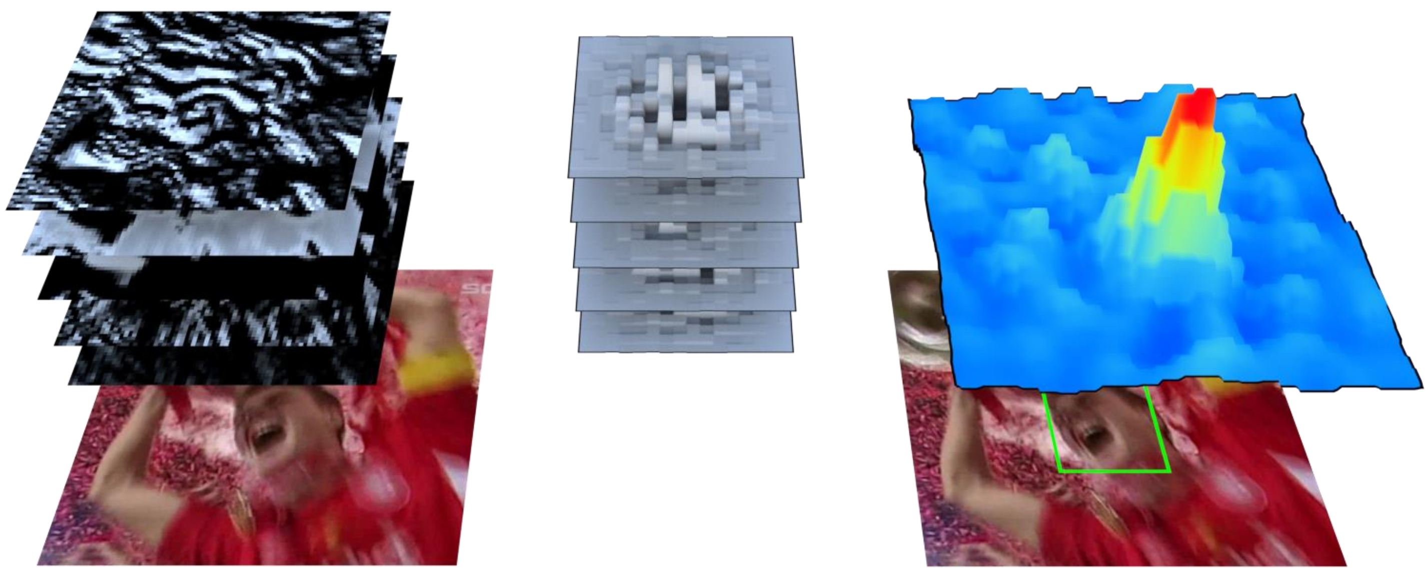
Limitations:

Single-resolution feature map

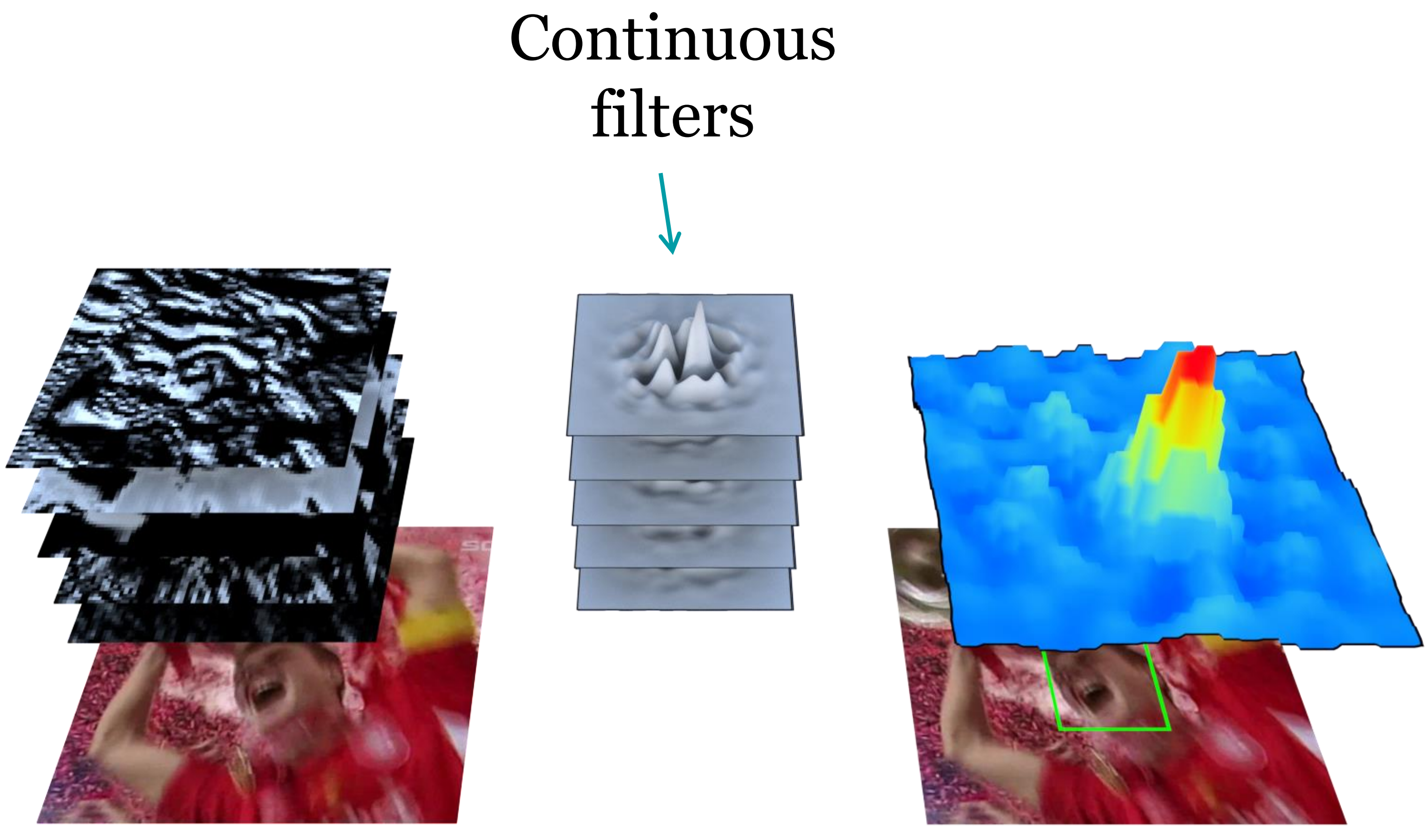
Coarse output scores



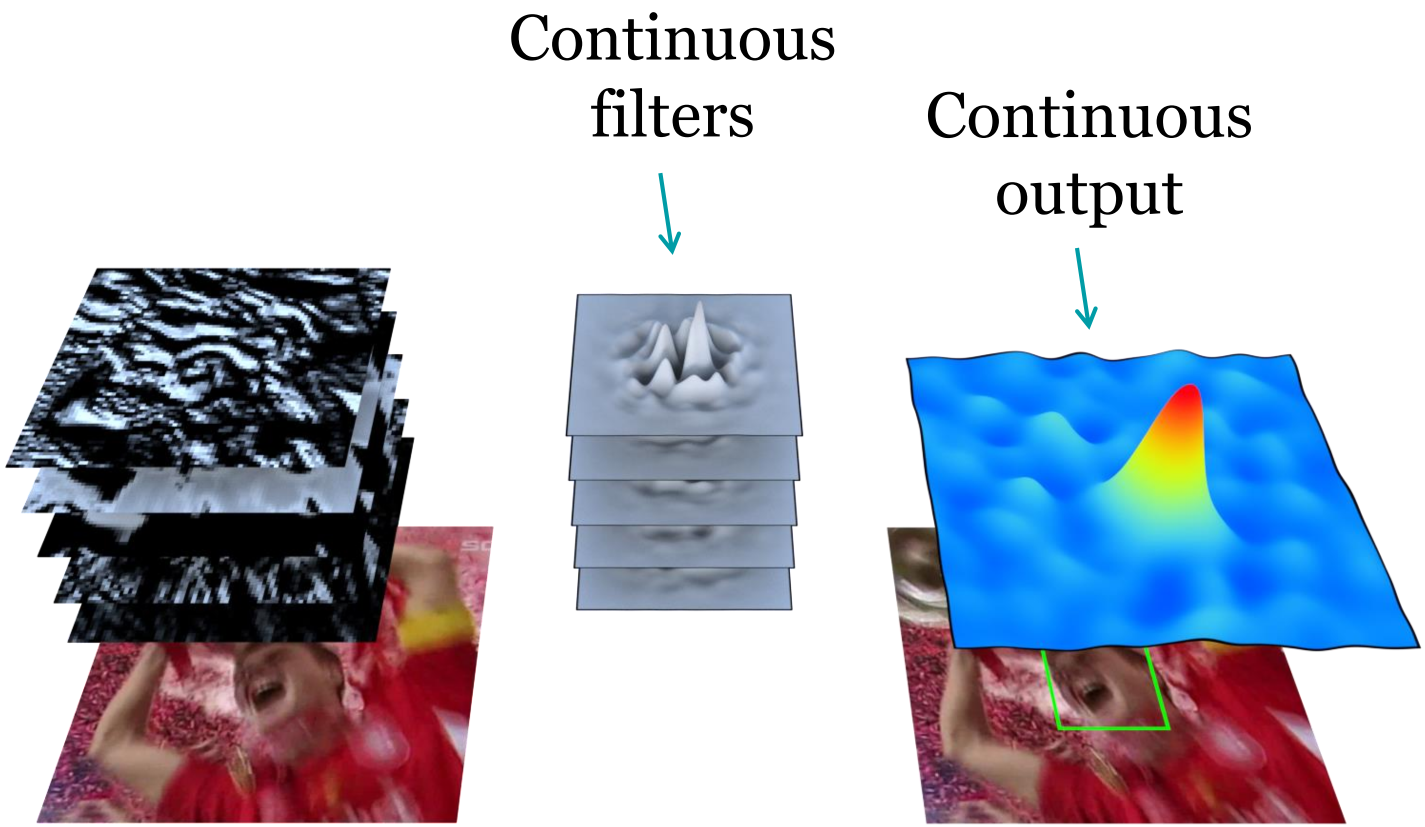
Our Approach: Overview



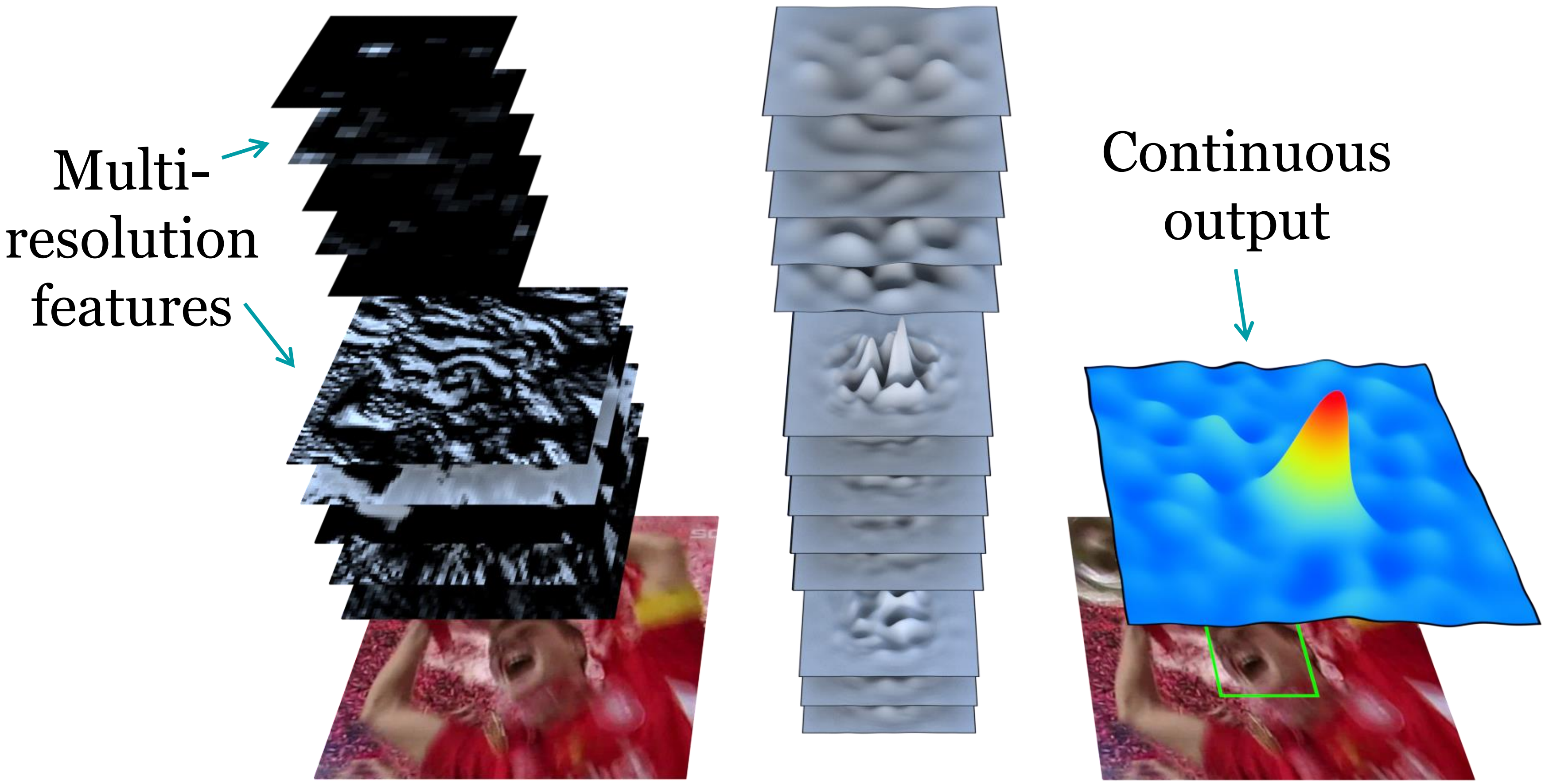
Our Approach: Overview



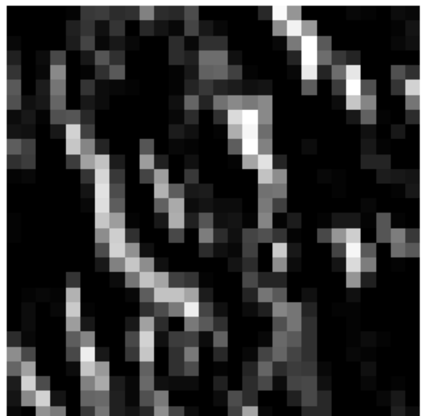
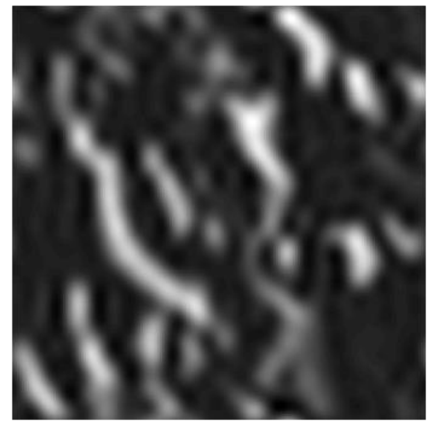
Our Approach: Overview

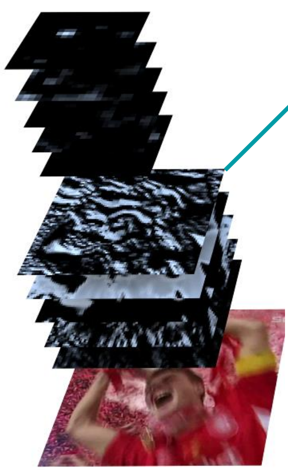


Our Approach: Overview



Interpolation Operator

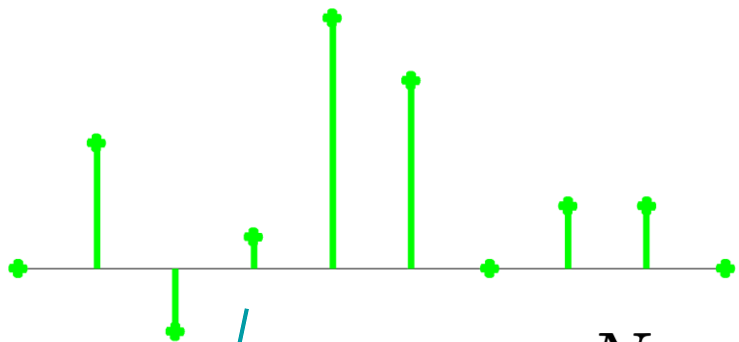
$J_d\{$  $\} =$ 



Interpolation Operator

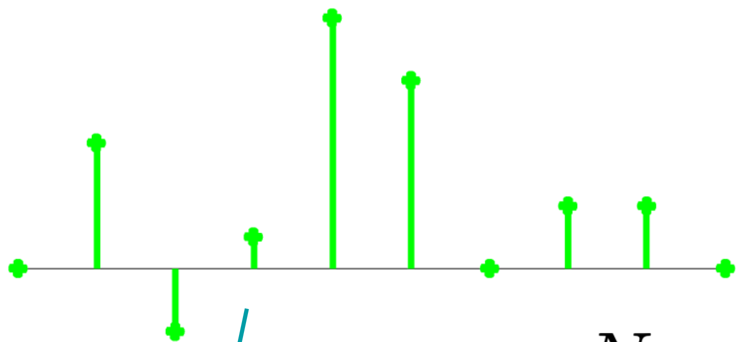
$$J_d\{x^d\}(t) = \sum_{n=0}^{N_d-1} x^d[n] b_d \left(t - \frac{T}{N_d} n \right)$$

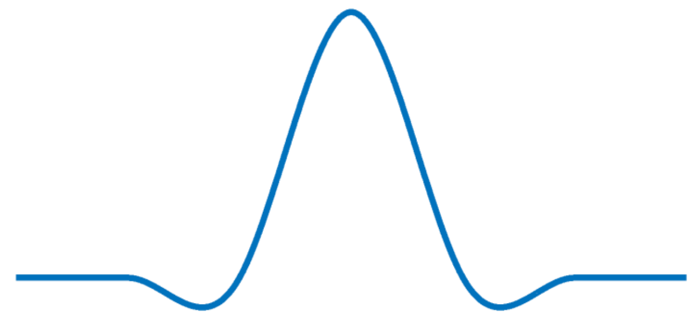
Interpolation Operator



$$J_d\{x^d\}(t) = \sum_{n=0}^{N_d-1} x^d[n] b_d \left(t - \frac{T}{N_d} n\right)$$

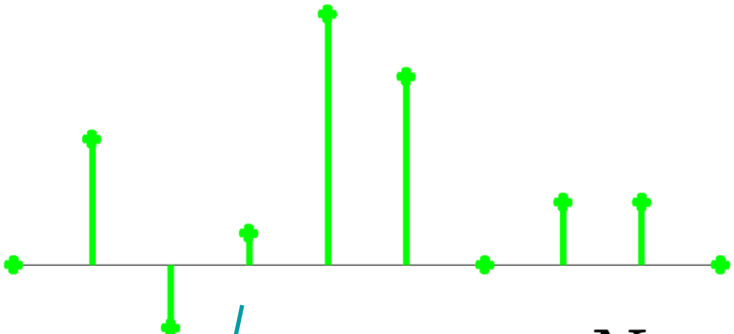
Interpolation Operator

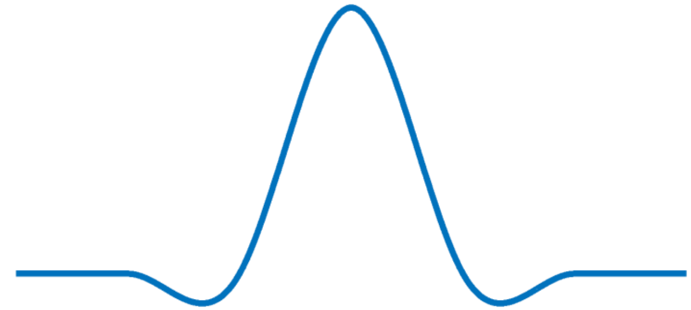




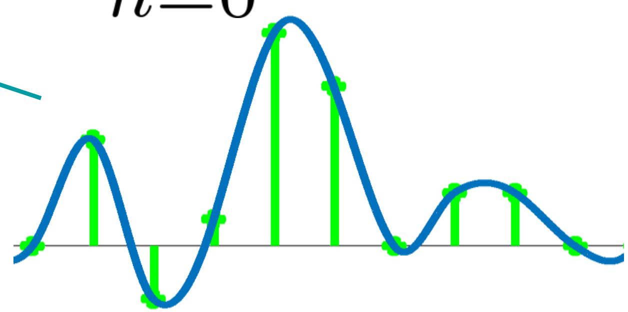
$$J_d\{x^d\}(t) = \sum_{n=0}^{N_d-1} x^d[n]\, b_d\left(t - \frac{T}{N_d}n\right)$$

Interpolation Operator

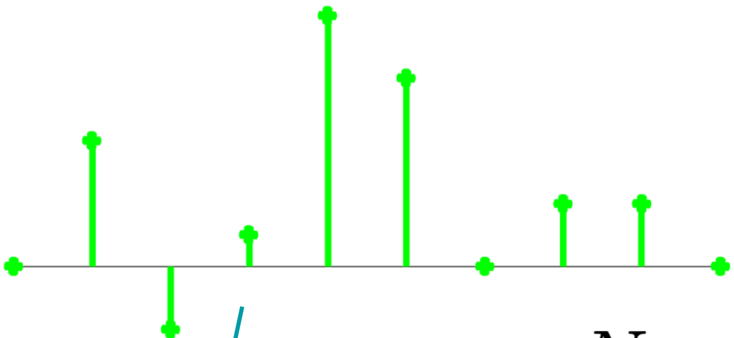


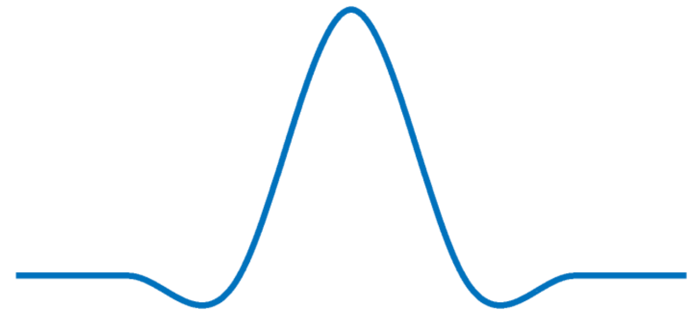


$$J_d\{x^d\}(t) = \sum_{n=0}^{N_d-1} x^d[n] b_d\left(t - \frac{T}{N_d}n\right)$$

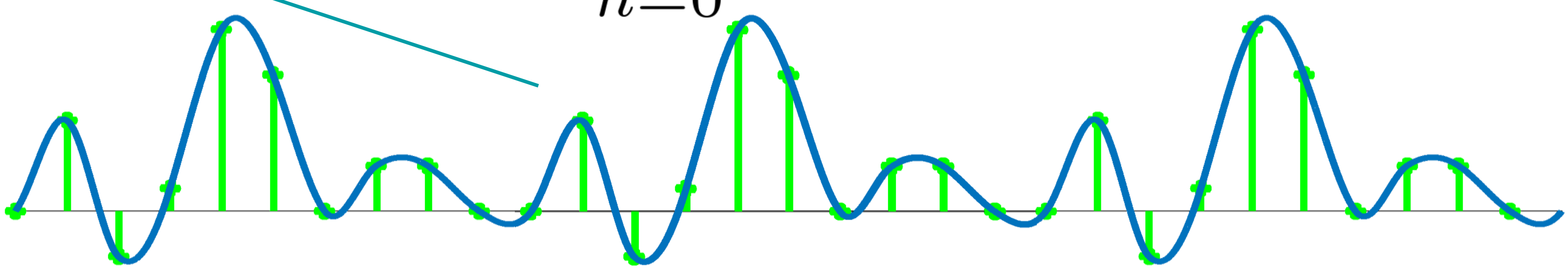


Interpolation Operator





$$J_d\{x^d\}(t) = \sum_{n=0}^{N_d-1} x^d[n]\, b_d\left(t - \frac{T}{N_d}n\right)$$



Convolution Operator

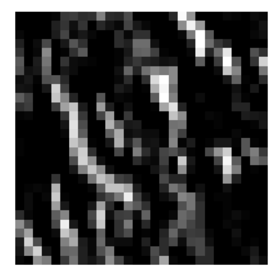
$$S_f\{x\} = \sum_{d=1}^{D} f^d * J_d\{x^d\}$$

Convolution Operator

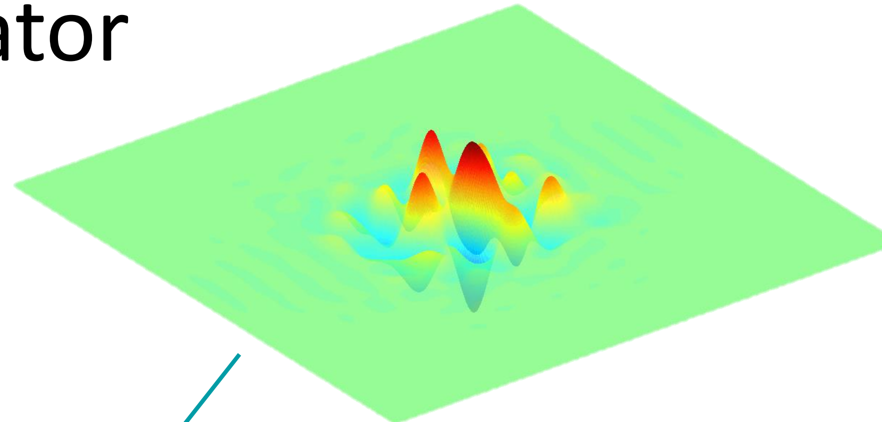
$$S_f\{x\} = \sum_{d=1}^{D} f^d * J_d\{x^d\}$$

Convolution Operator

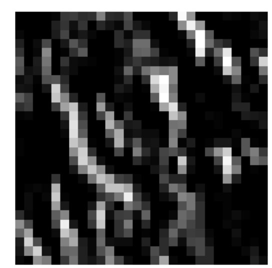
$$S_f\{x\} = \sum_{d=1}^{D} f^d * J_d\{x^d\}$$



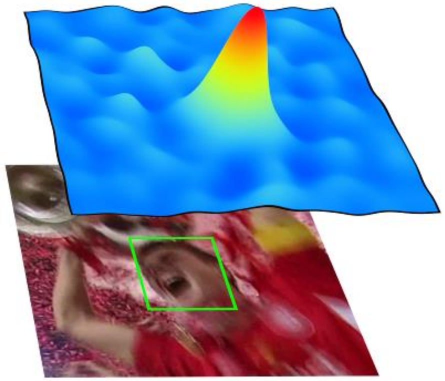
Convolution Operator

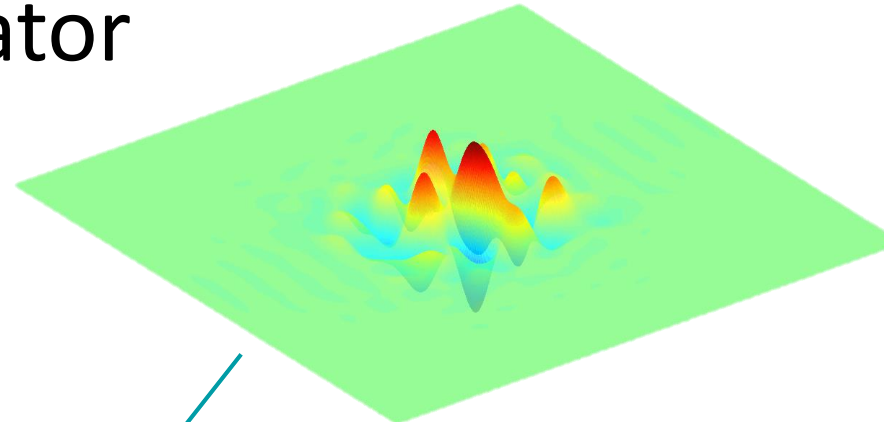


$$S_f\{x\} = \sum_{d=1}^{D} f^d * J_d\{x^d\}$$

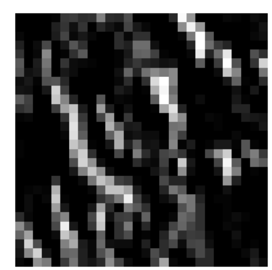


Convolution Operator

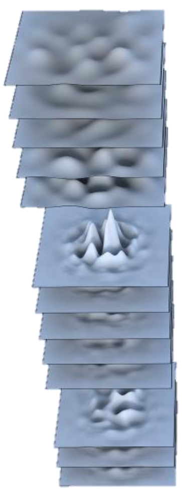




$$S_f\{x\} = \sum_{d=1}^{D} f^d * J_d\{x^d\}$$

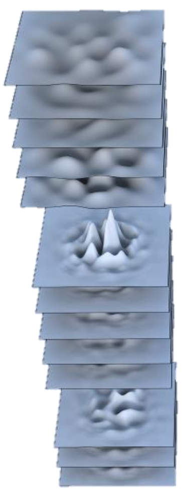


Training Loss



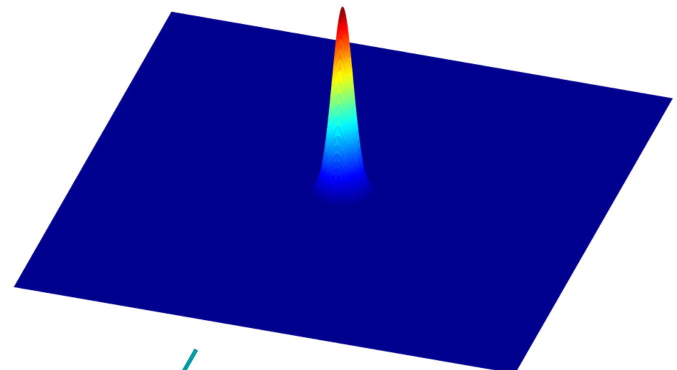
$$E(f) = \sum_{j=1}^{m} \alpha_j \left\| S_f\{x_j\} - y_j \right\|^2 + R(f)$$

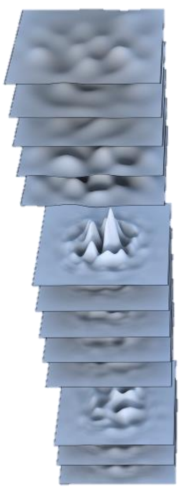
Training Loss



$$E(f) = \sum_{j=1}^{m} \alpha_j \left\| S_f\{x_j\} - y_j \right\|^2 + R(f)$$

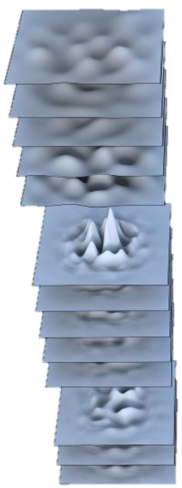
Training Loss

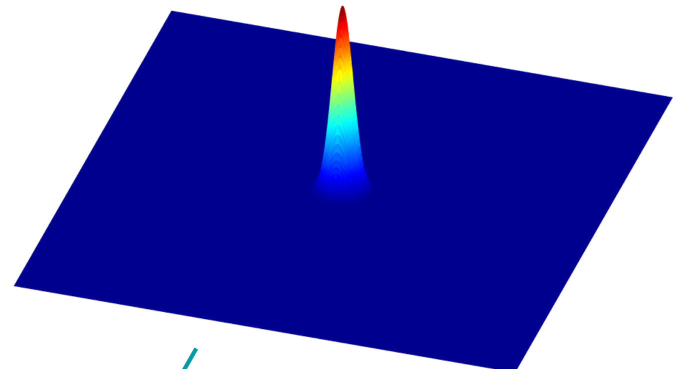




$$E(f) = \sum_{j=1}^{m} \alpha_j \left\| S_f\{x_j\} - y_j \right\|^2 + R(f)$$

Training Loss





$$E(f) = \sum_{j=1}^{m} \alpha_j \left\| S_f\{x_j\} - y_j \right\|^2 + \underbrace{R(f)}_{}$$

[Danelljan et al., ICCV 2015]

Localization

$$S_f\{ \qquad\qquad\qquad \} =$$

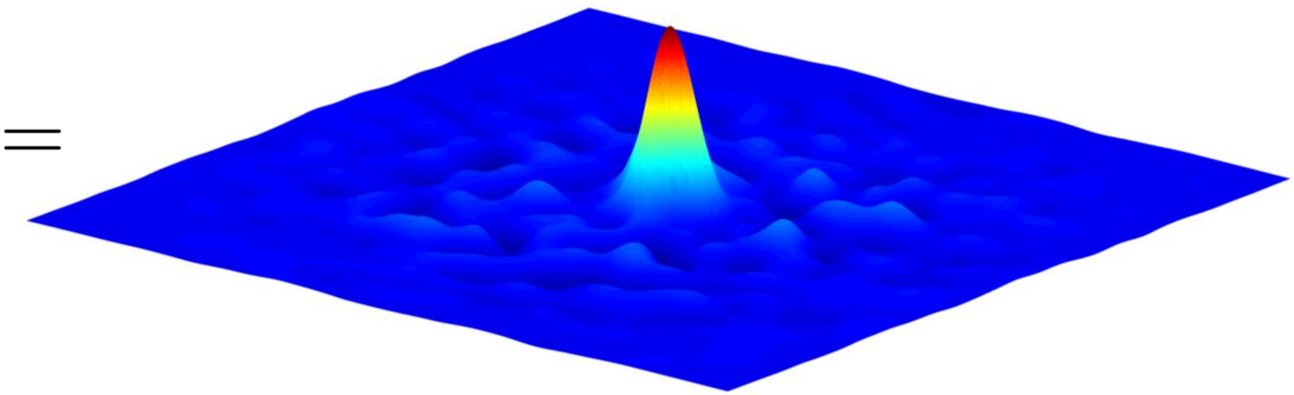
Localization

$S_f\{$  $\} =$

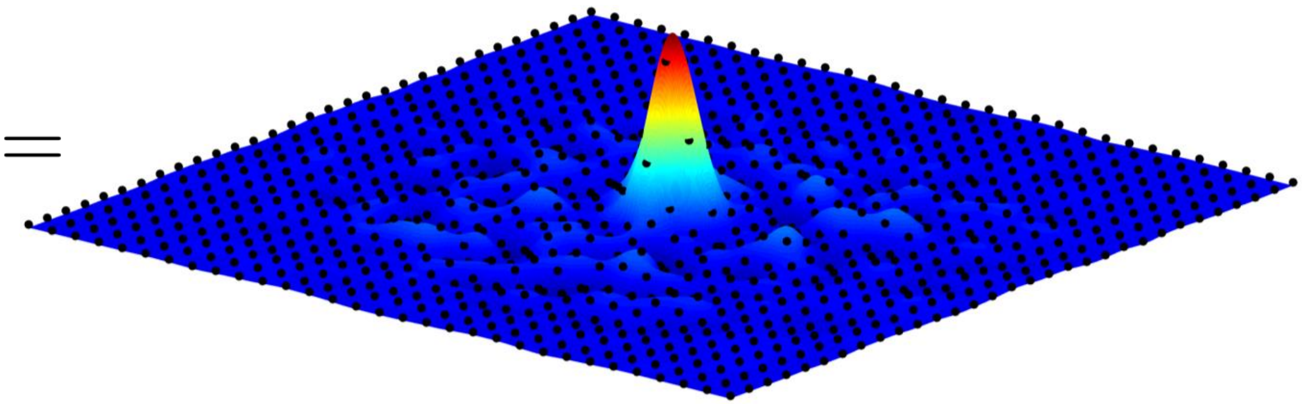
Localization

$S_f\{$

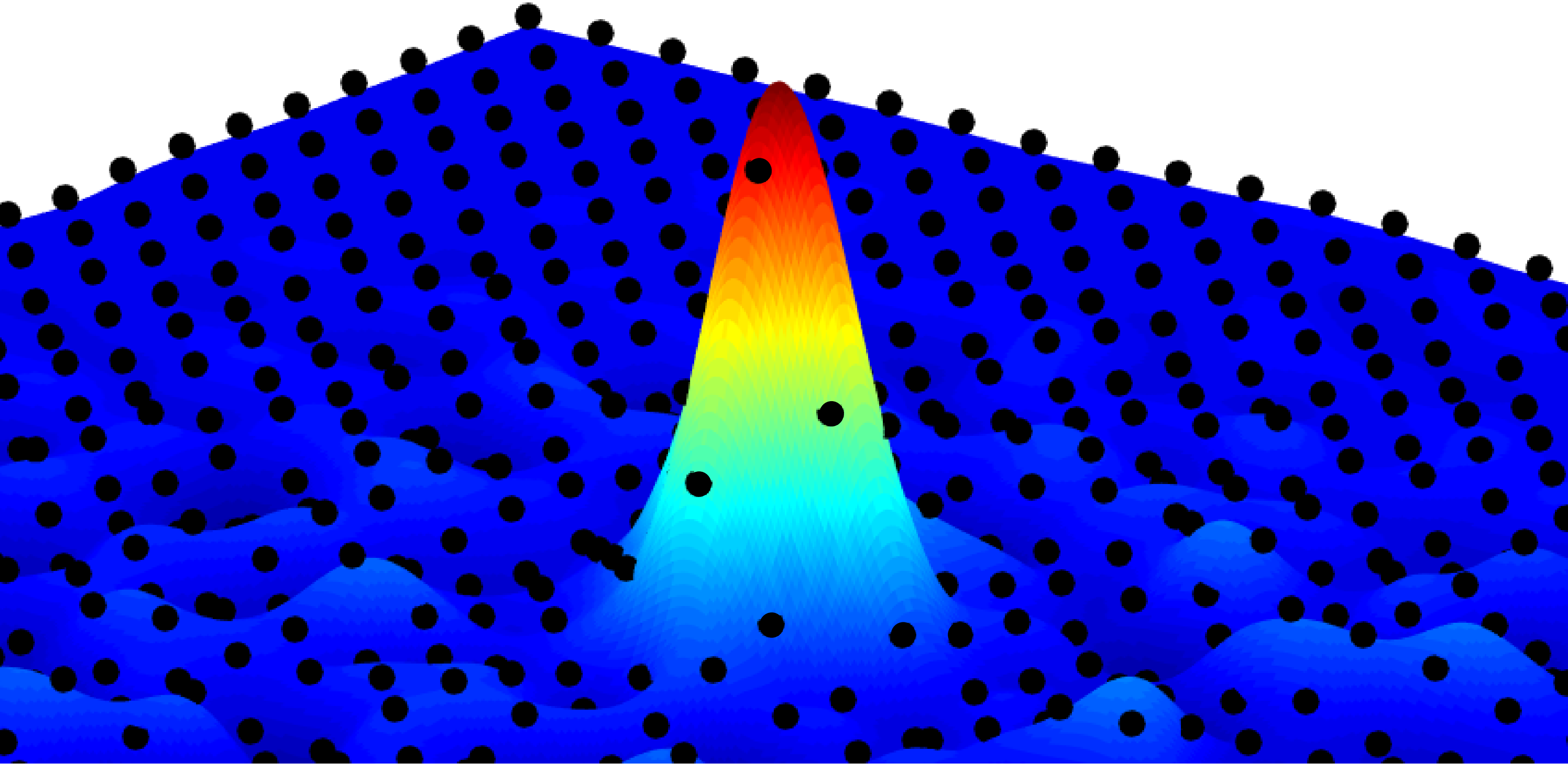
$\} =$



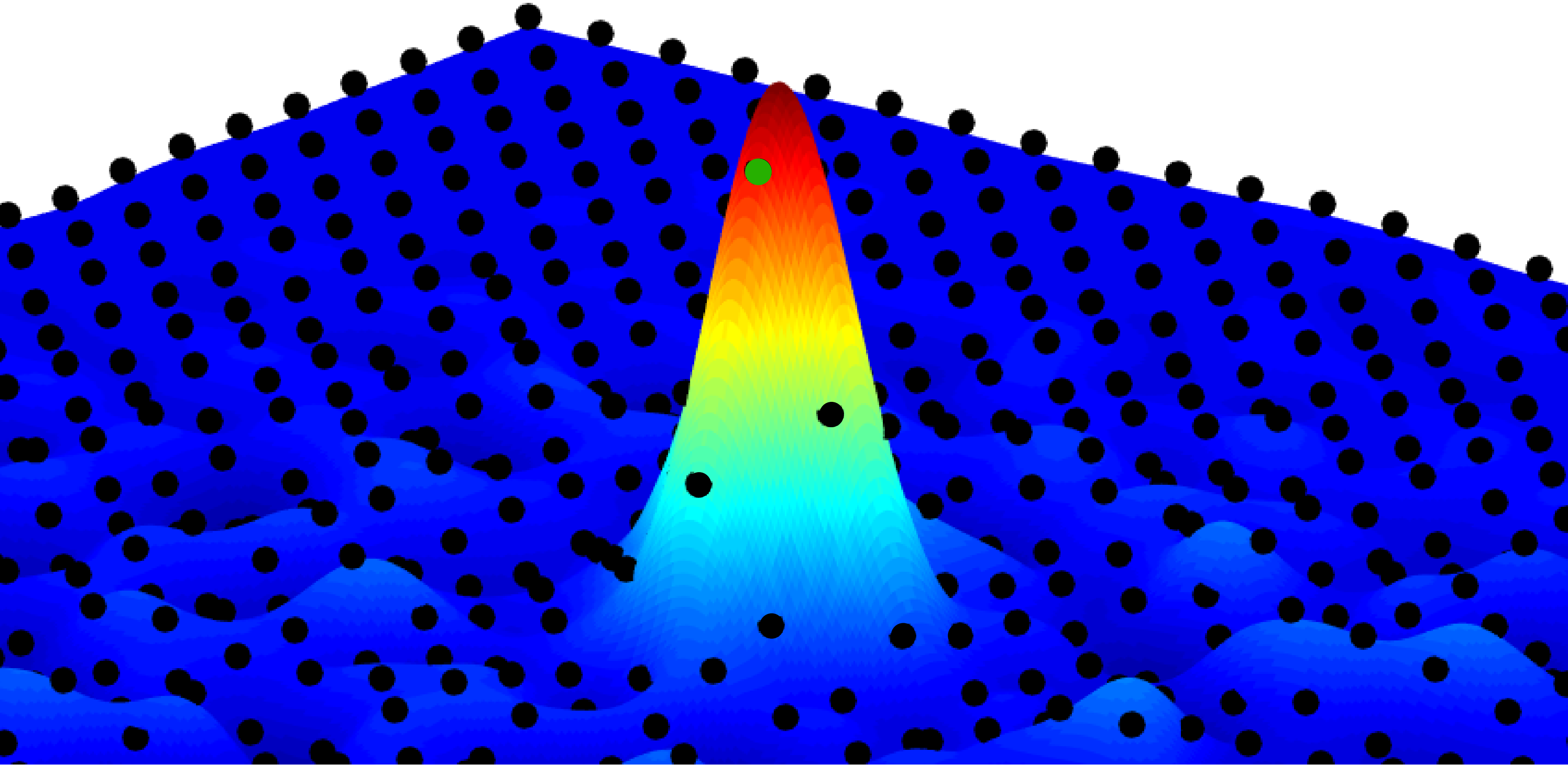
Localization

$S_f\{$  $\} =$ 

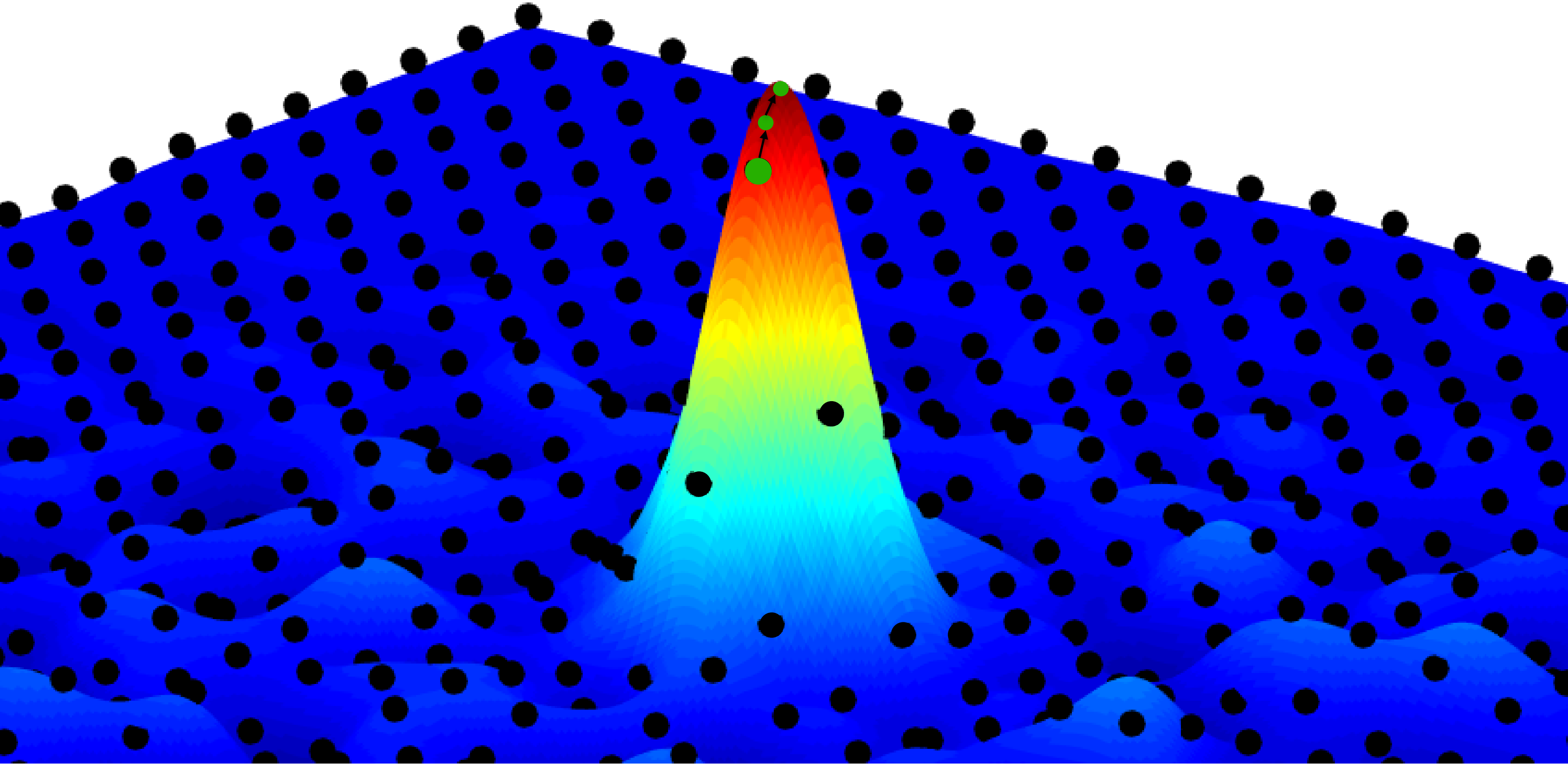
Localization



Localization



Localization



Object Tracking Framework

- Features: VGG
  - Pre-trained on ImageNet
  - No fine-tuning on application specific data

Object Tracking Framework

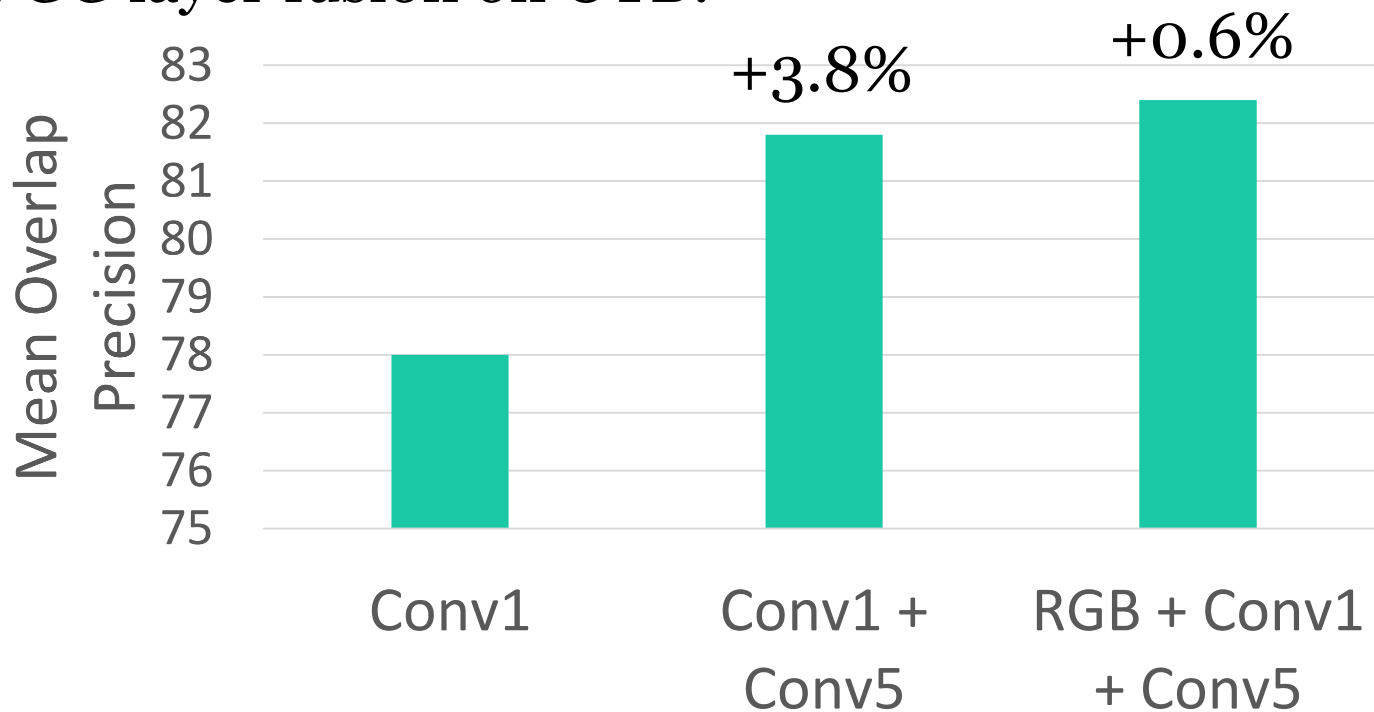
- Features: VGG
  - Pre-trained on ImageNet
  - No fine-tuning on application specific data
- Optimization: Conjugate Gradient

Experiments: Object Tracking

- 3 datasets: OTB-100, Temple-Color, VOT2015

Experiments: Object Tracking

- 3 datasets: OTB-100, Temple-Color, VOT2015
- VGG layer fusion on OTB:

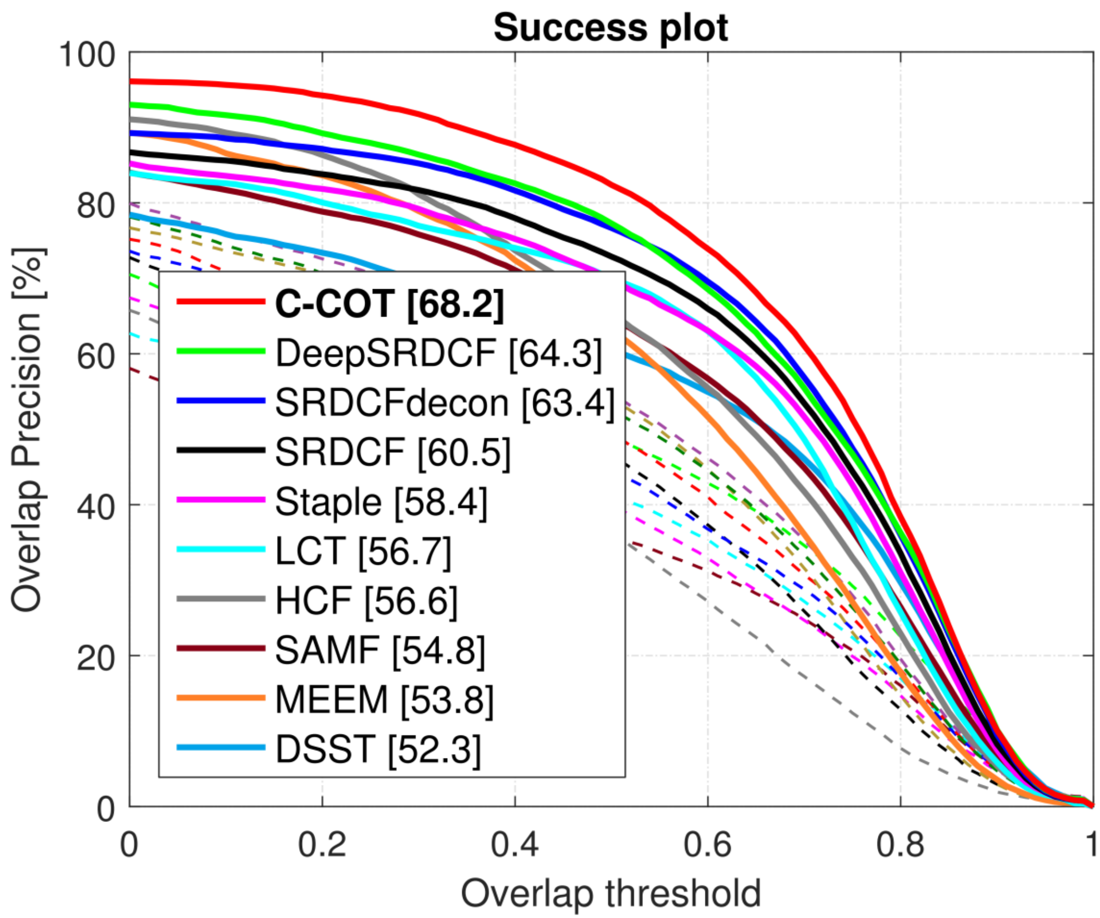


Experiments: Object Tracking

- Compared to explicit resampling in DCF
  - Performance gain: $+7.4\%$ AUC
  - Efficiency gain: -80% data size

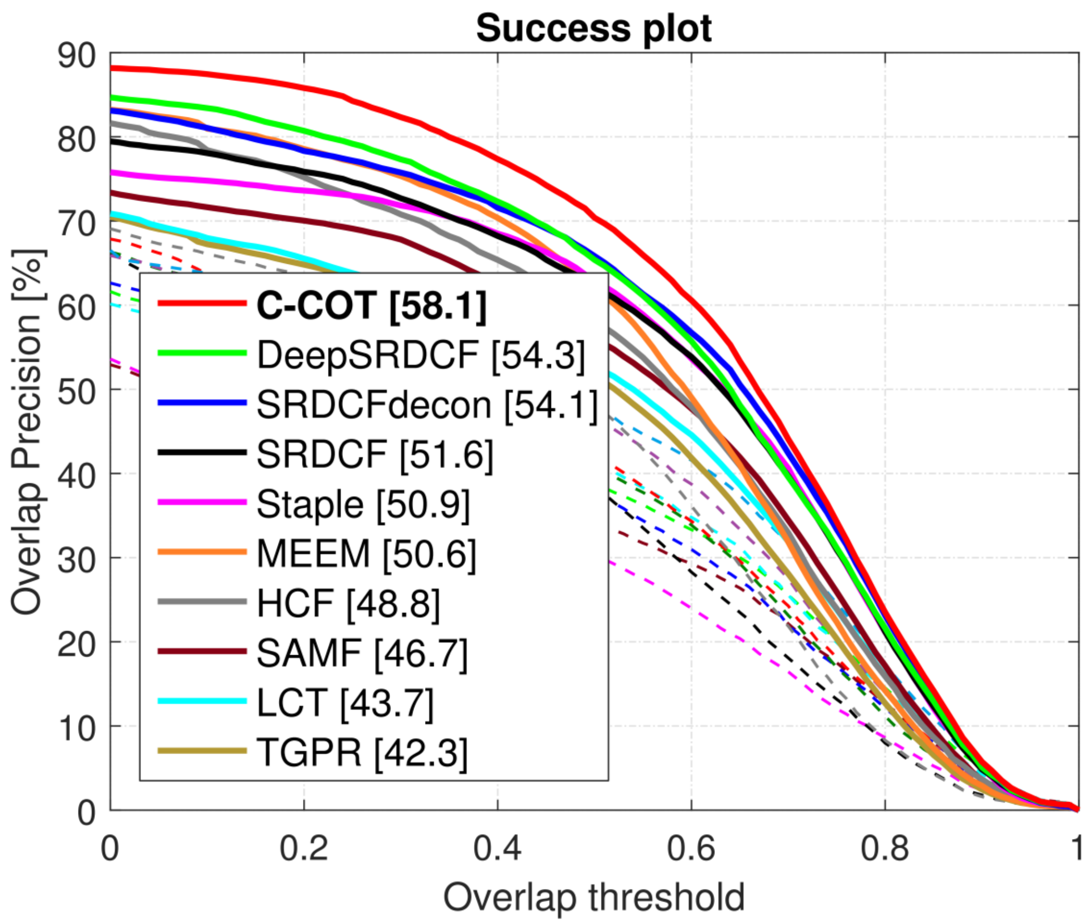
OTB Dataset (100 videos)

+3.9% AUC



Temple-Color Dataset (128 videos)

+3.8% AUC



Visual Object Tracking Challenge 2016

	Tracker	EAO	A	R	A_{rank}	R_{rank}	AO	EFO	Impl.
1.	C-COT	**0.331**	0.539	*0.238*	12.000	**1.000**	0.469	0.507	D M
2.	TCNN	*0.325*	0.554	0.268	4.000	*2.000*	*0.485*	1.049	S M
3.	SSAT	0.321	**0.577**	0.291	**1.000**	3.000	**0.515**	0.475	S M
4.	MLDF	0.311	0.490	**0.233**	36.000	**1.000**	0.428	1.483	D M
5.	Staple	0.295	0.544	0.378	5.000	10.000	0.388	11.144	D C

[Matej et al., ECCV VOT workshop 2016]

Feature Point Tracking Framework

- Image intensity features
- Uniform regularization

Feature Point Tracking Framework

- Image intensity features
- Uniform regularization

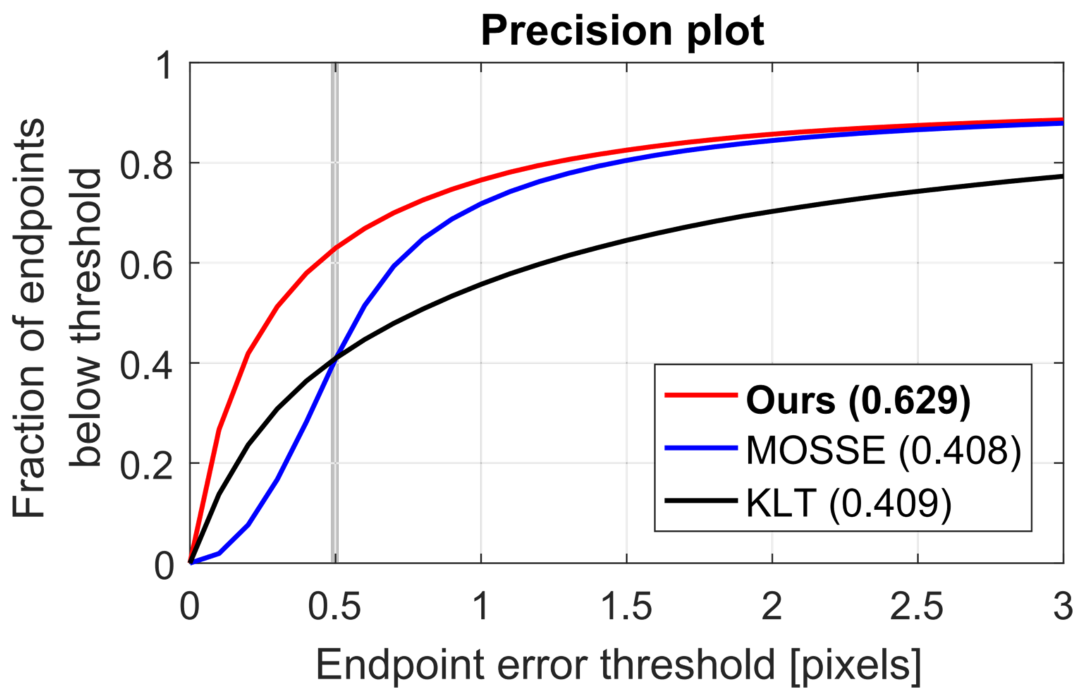
$$\Downarrow$$

$$\hat{f}[k] = \frac{\sum_{j=1}^{m} \alpha_j \overline{X_j[k]\hat{b}[k]}\hat{y}_j[k]}{\sum_{j=1}^{m} \alpha_j \left|X_j[k]\hat{b}[k]\right|^2 + \beta^2}$$

Experiments: Feature Point Tracking

- Dataset: Sintel

+22.1% inliers
at 0.5 pixels



Conclusions

- Learn **Continuous Convolution Operators**
  - **Multi-resolution** deep feature maps
  - **Sub-pixel** accurate localization
  - **Sub-pixel** supervision
- Superior results for two applications
  - Object tracking
  - Feature point tracking

Martin Danelljan

www.liu.se

LINKÖPING UNIVERSITY