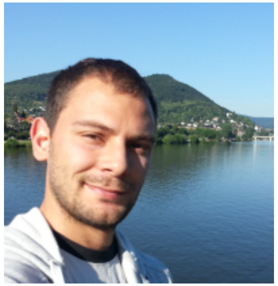
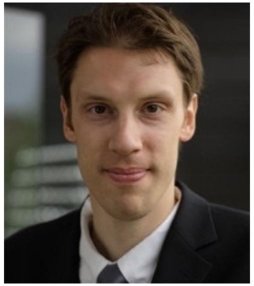
Cross and Learn:
Cross-Modal Self-Supervision

Nawid Sayed

Biagio Brattoli



Prof. Björn Ommer



HCI / IWR, Heidelberg University

Heidelberg, Germany

Motivation

Many real-world applications utilize supervised pre-training

Object detection [1]

Semantic segmentation [2]

Image captioning [3]

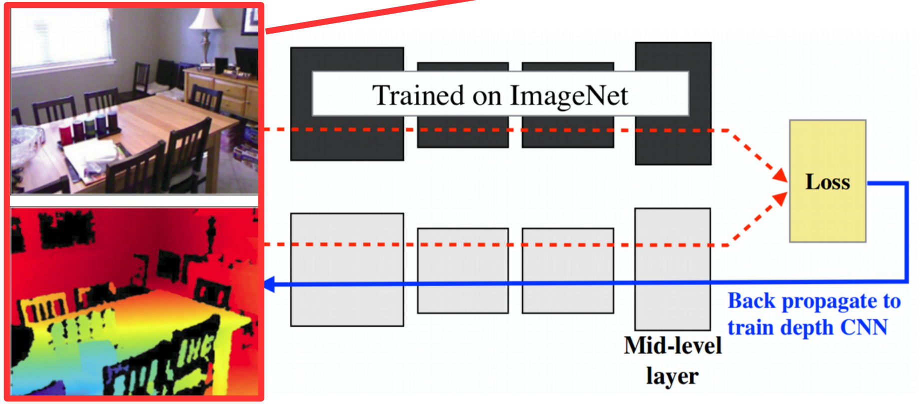
[1] Redmon et al., “You Only Look Once: Unified, Real-Time Object Detection” (2015)

[2] Zhao et al., “Pyramid Scene Parsing Network” (2016)

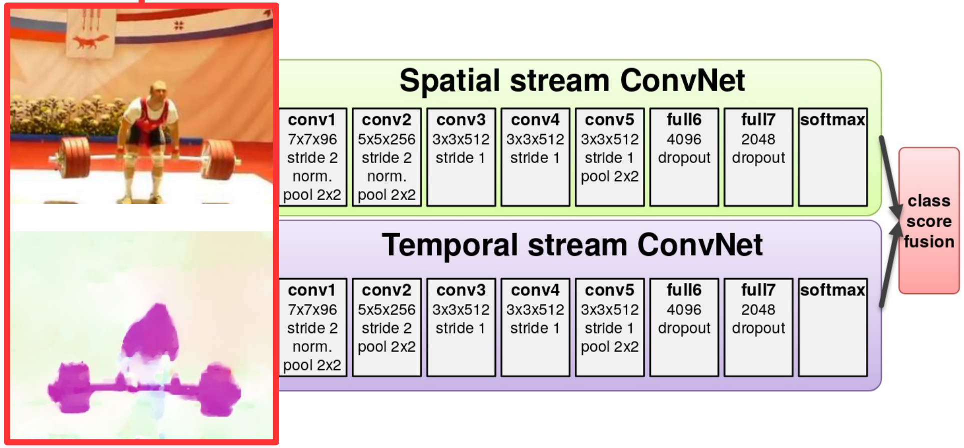
[3] Vinyals et al., “Show and Tell: A Neural Image Caption Generator” (2014)

Motivation

Supervised learning from paired multi-modal data successful



Knowledge transfer from RGB to depth modality [1]



Action recognition jointly using RGB and optical flow [2]

[1] Gupta et al., “Cross-Modal Distillation for Supervision Transfer” (2015)

[2] Simonyan et al., “Two-Stream Convolutional Networks for Action Recognition in Videos” (2014)

Motivation

Large scale unannotated image and video data free

Motivation

Large scale unannotated image and video data free

Supervised learning relies on annotated data

Motivation

Large scale unannotated image and video data free

Supervised learning relies on annotated data

Annotations costly and prone to errors

“Siberian Husky” “Eskimo Dog”

Samples from the ImageNet dataset [1]

[1] Russakovsky et al., “ImageNet Large Scale Visual Recognition Challenge” (2014)

Motivation

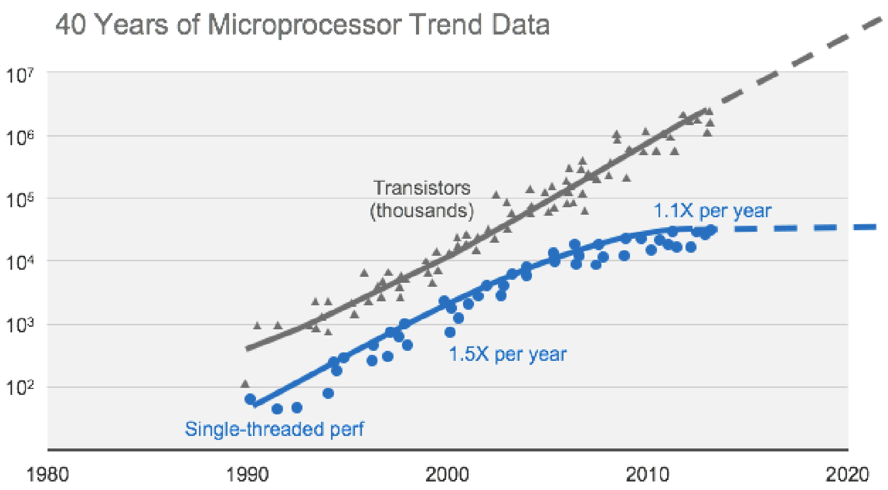
Large scale unannotated image and video data free

Supervised learning relies on annotated data

Annotations costly and prone to errors

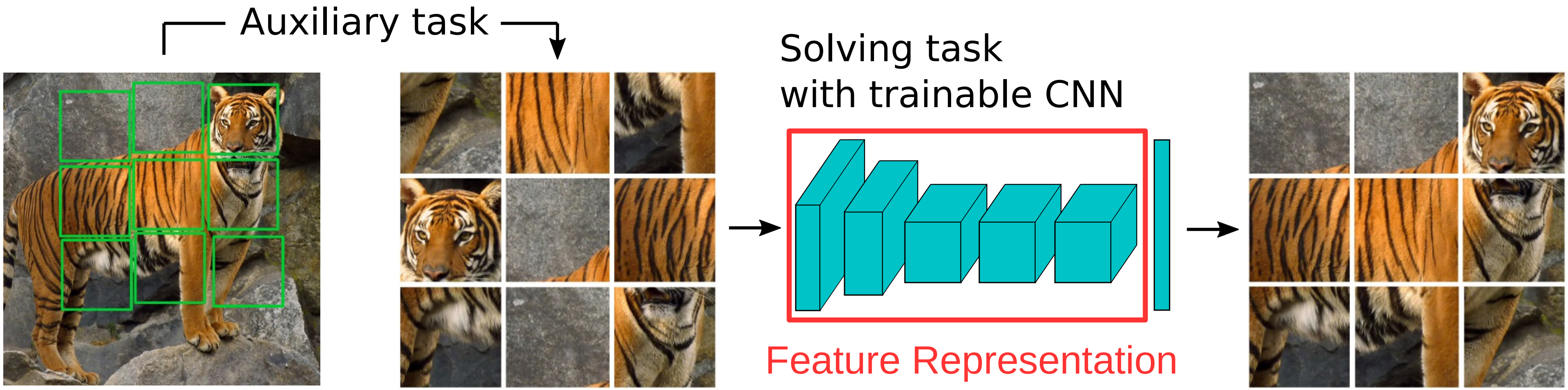
Supervised learning does not scale well into future

Increase in computational power



Motivation

Self-supervised representation learning



Solving Jigsaw Puzzles [1]

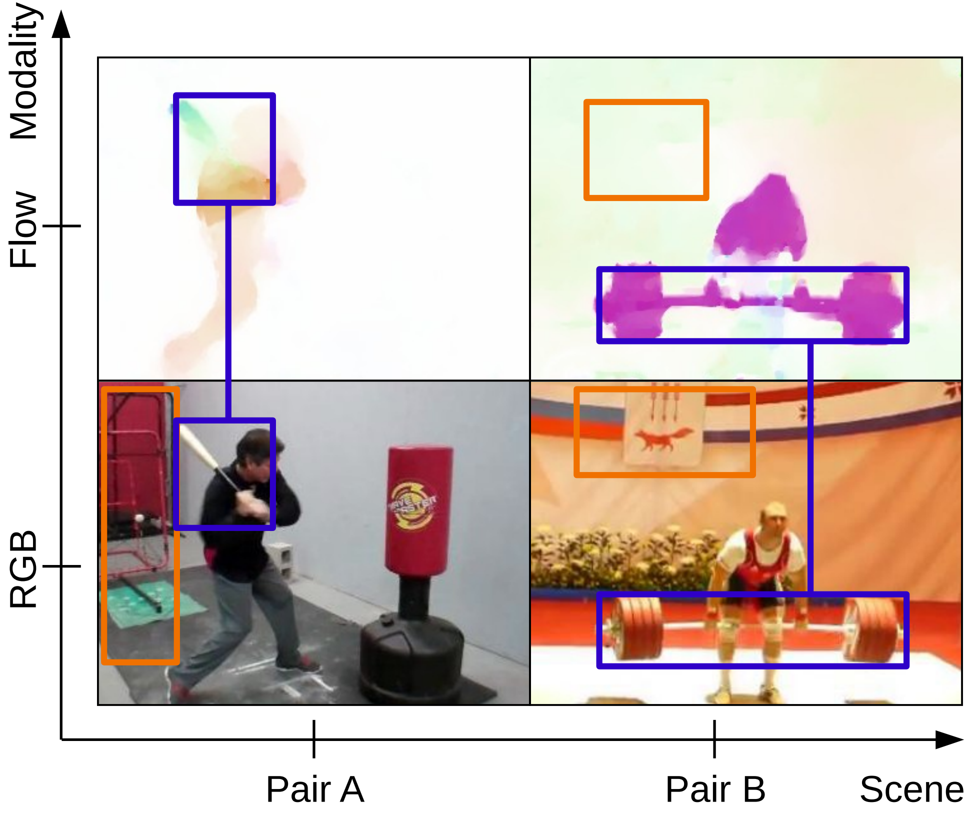
[1] Noroozi et al., “Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles” (2016)

Approach

Self-supervised learning from paired multi-modal data

Cross-modal information has high semantic meaning (barbell, bat)

Modalitiy specific content has low semantic meaning (background, camera motion)



Approach

Desirable features:

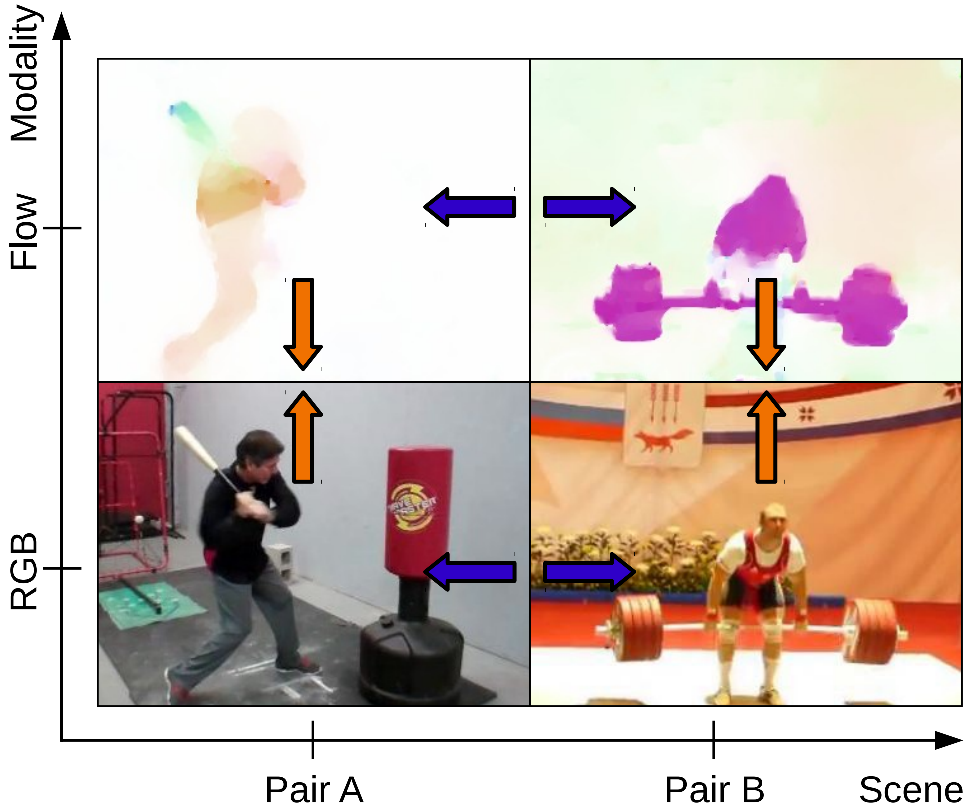
- Invariant to modality specific content

  Similar features in a pair

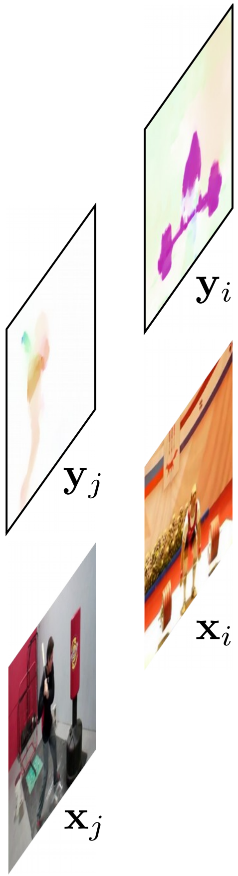
- Sensitive to cross-modal information

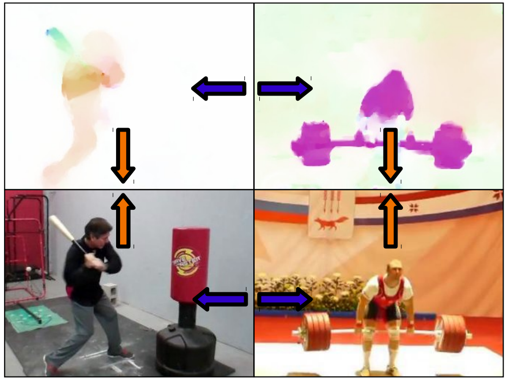
  Distant features between pairs

Achieved using L_{cross} and L_{div}

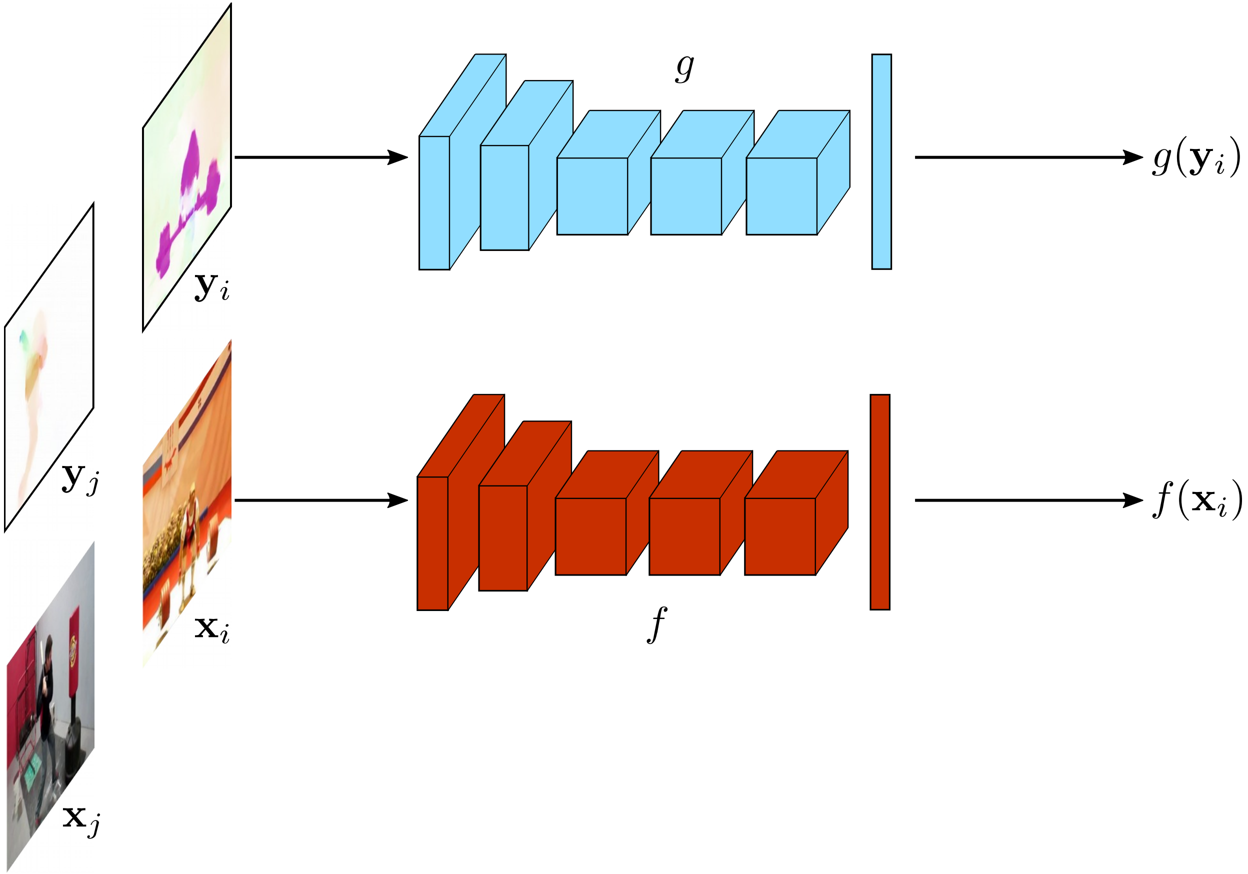


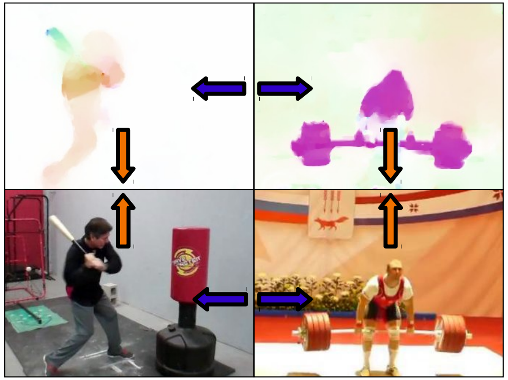
Model Pipeline



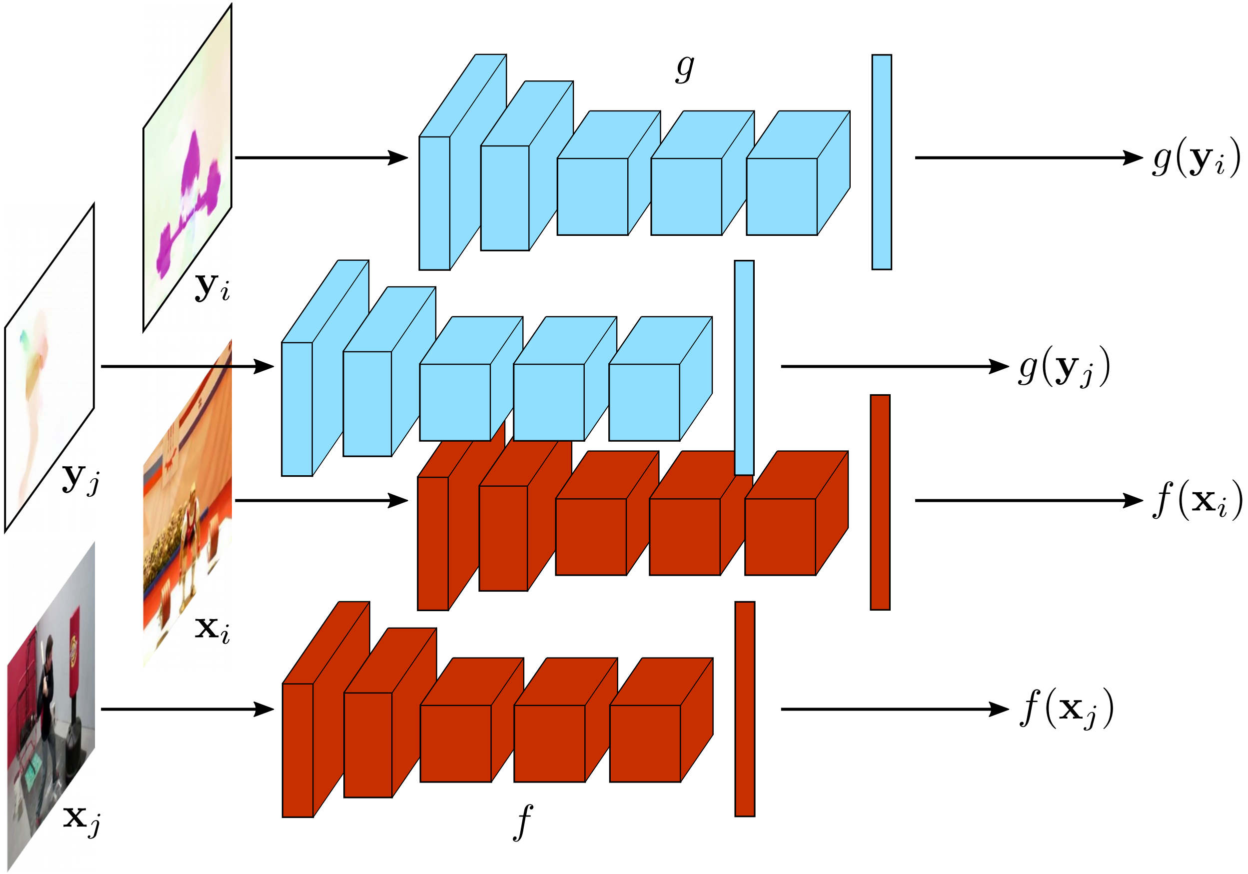


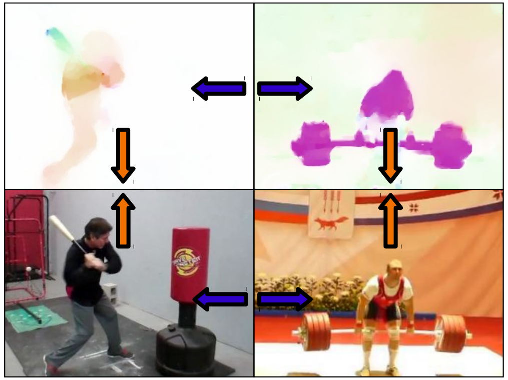
Model Pipeline



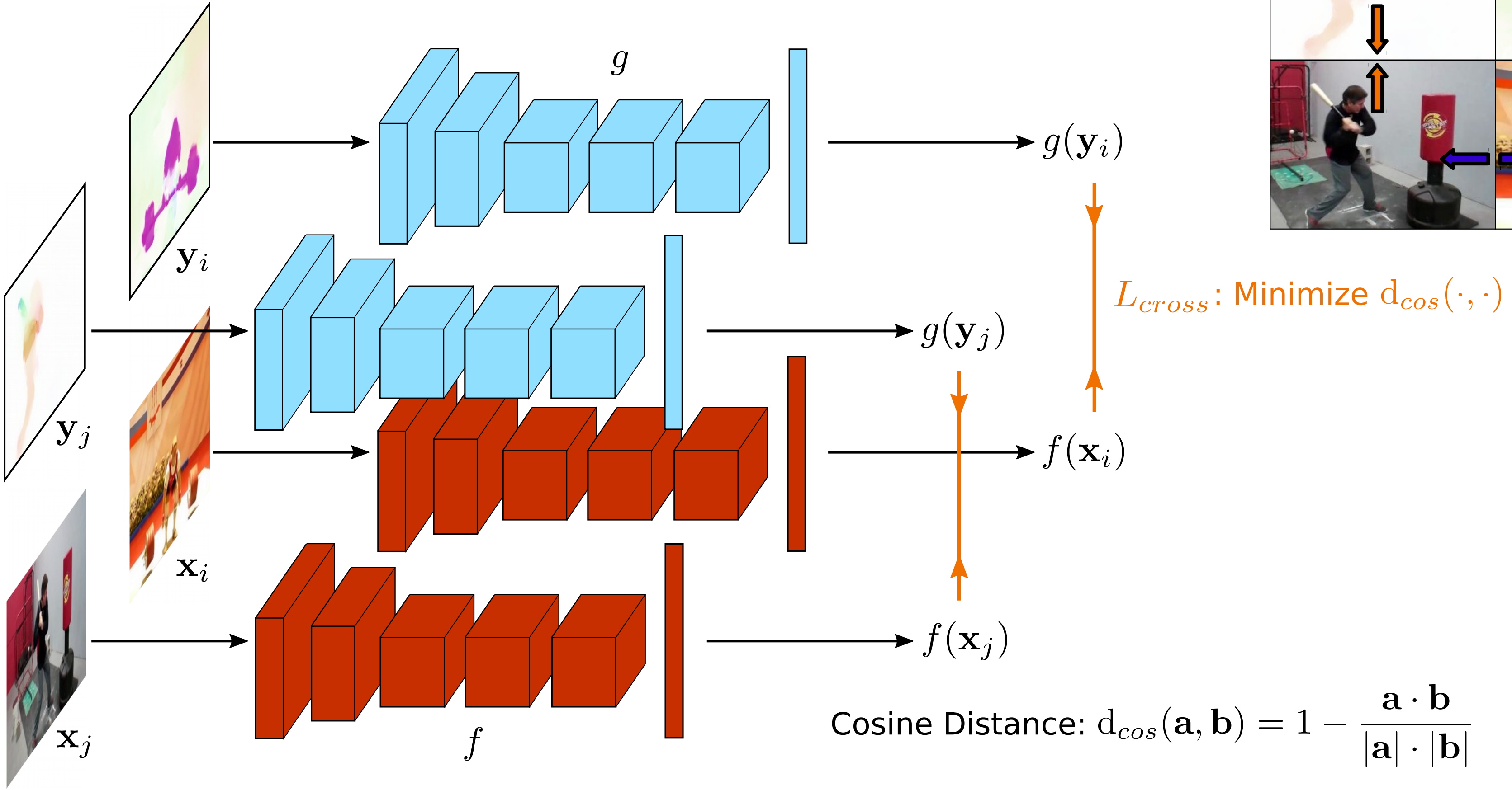


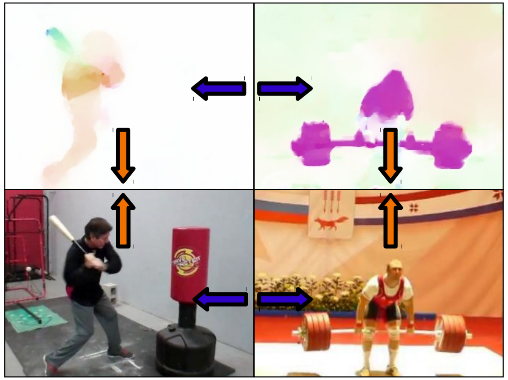
Model Pipeline





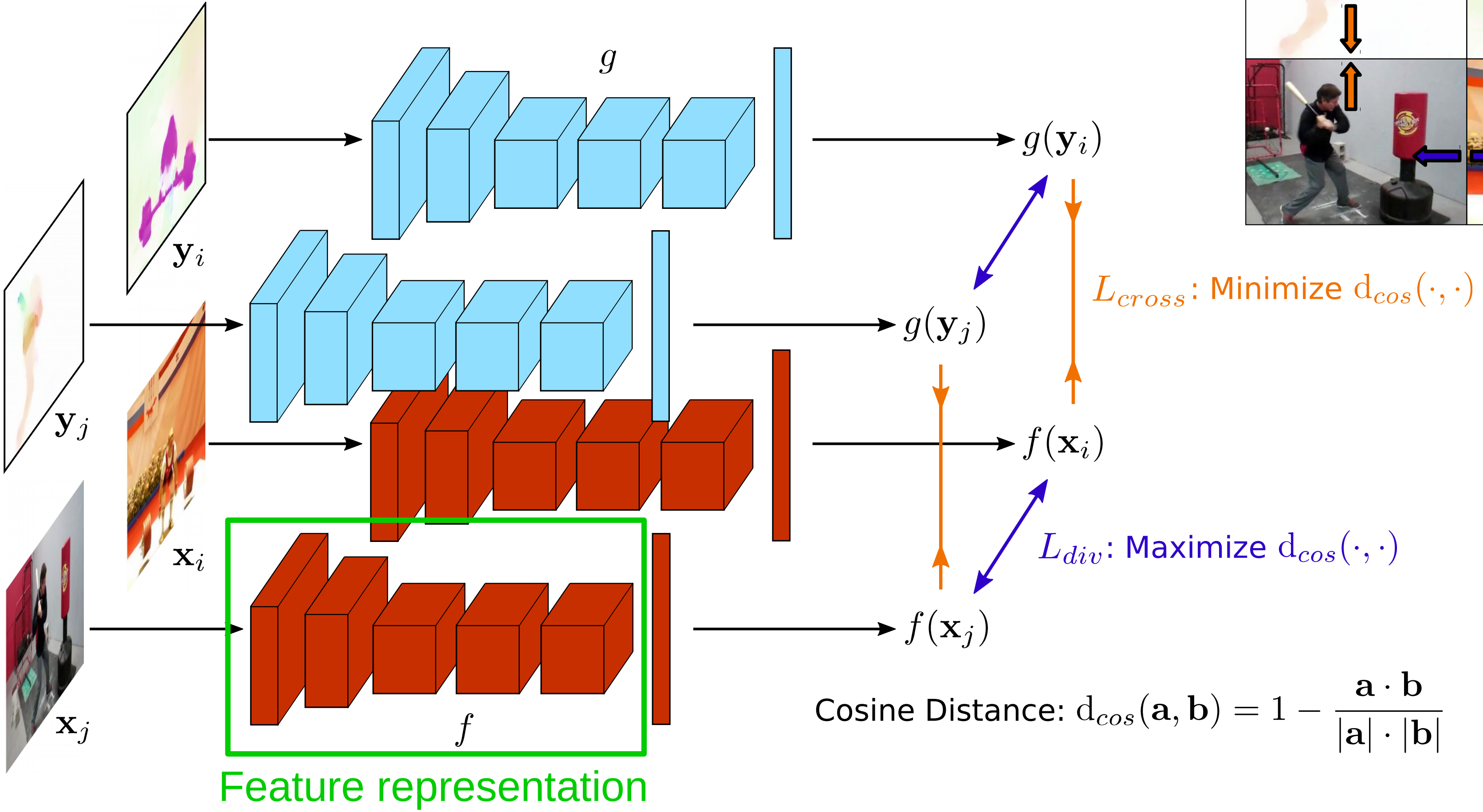
Model Pipeline

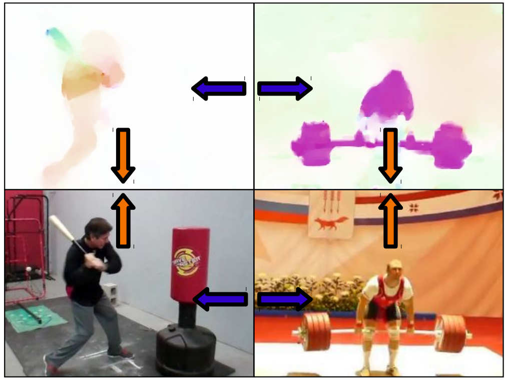




Cosine Distance: $\mathrm{d}_{cos}(\mathbf{a},\mathbf{b}) = 1 - \dfrac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{a}|\cdot|\mathbf{b}|}$

Model Pipeline

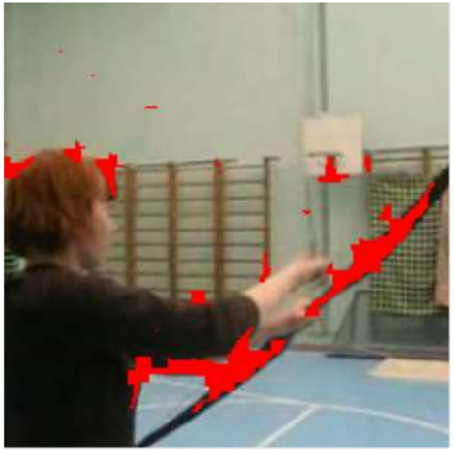




Cosine Distance: $\mathrm{d}_{cos}(\mathbf{a},\mathbf{b}) = 1 - \frac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{a}|\cdot|\mathbf{b}|}$

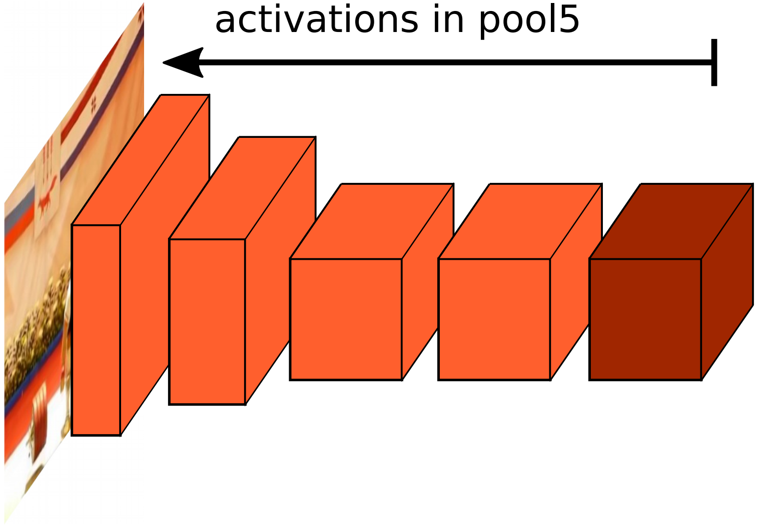
High-Level Input Activations

Input

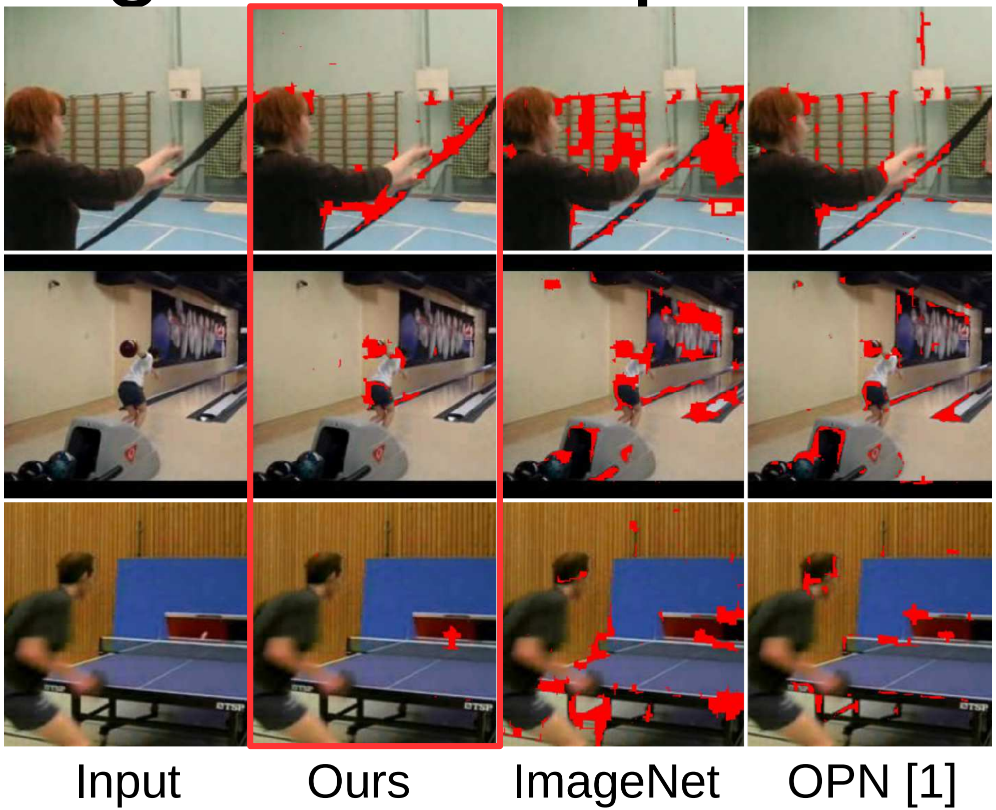


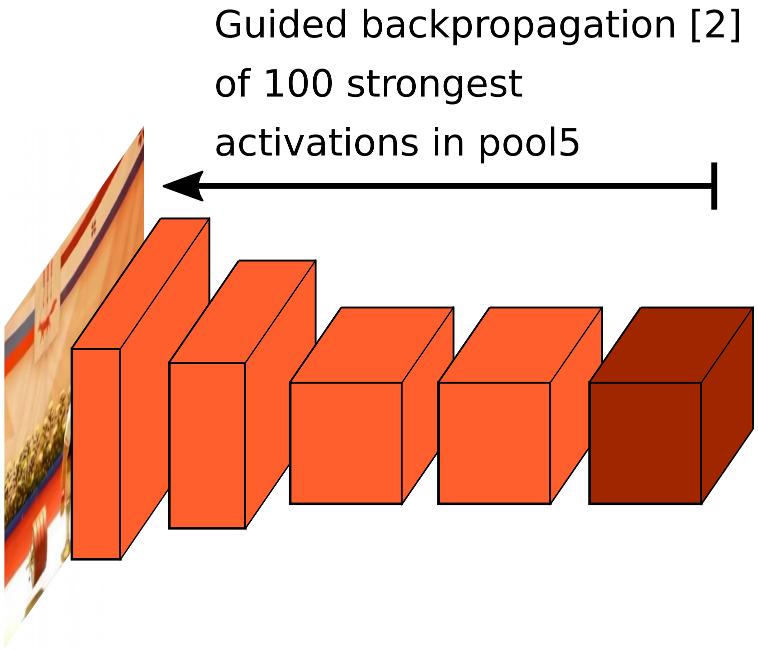
regions of interest

Guided backpropagation [2] of 100 strongest activations in pool5



High-Level Input Activations

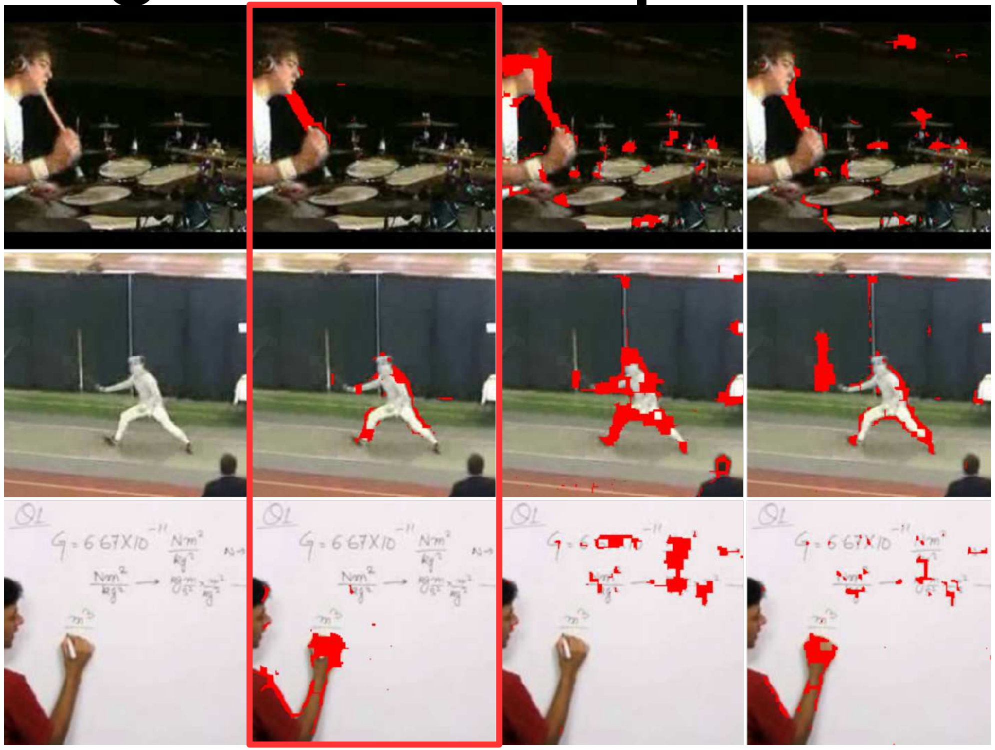




[1] Lee et al., “Unsupervised Representation Learning by Sorting Sequences” (2017)

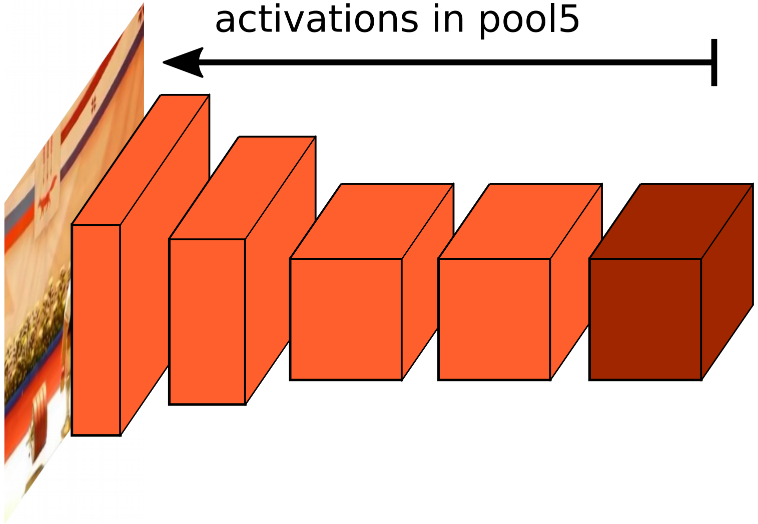
[2] Springenberg et al., “Striving for Simplicity: The All Convolutional Net” (2014)

High-Level Input Activations



Input Ours ImageNet OPN [1]

Guided backpropagation [2] of 100 strongest activations in pool5



[1] Lee et al., “Unsupervised Representation Learning by Sorting Sequences” (2017)

[2] Springenberg et al., “Striving for Simplicity: The All Convolutional Net” (2014)

Action Recognition

	Pre-training data	Traintime	UCF-101	HMDB-51
Random	None	None	48.2	19.5
ImageNet [5]	ImageNet	3 days	67.7	28.0
Shuffle and Learn [1]	UCF-101	-	50.2	18.1
VGAN [2] (C3D)	flickr (2M videos)	$>$ 2 days	52.1	-
LT-Motion[3] (RNN)	NTU (57K videos)	-	53.0	-
Pose f. Action [4] (VGG)	UCF,HMDB,ACT	-	55.0	23.6
OPN [5]	UCF-101	40 hours	56.3	22.1
Our	UCF-101	**6 hours**	**58.7**	**27.2**
Random (VGG16)+	None	None	59.6	24.3
Our (VGG16)+	UCF-101	1.5 days	70.5	33.0

[1] Misra et al., “Shuffle and Learn Unsupervised Learning using Temporal Order Verification” (2016)

[2] Vondrick et al., “Generating Videos with Scene Dynamics” (2016)

[3] Luo et al., “Unsupervised Learning of Long-Term Motion Dynamics for Videos” (2017)

[4] Purushwalkam et al., “Pose from Action: Unsupervised Learningof Pose Features based on Motion” (2016)

[5] Lee et al., “Unsupervised Representation Learning by Sorting Sequences” (2017)

Transfer Learning

Pascal VOC 2007 object classification and detecion

	Pre-training data	Traintime	Classification	Detection
ImageNet [5]	ImageNet	3 days	78.2	56.8
Context [1]	ImageNet	4 weeks	55.3	46.6
Counting [2]	ImageNet	-	67.7	51.4
Jigsaw [3]	ImageNet	2.5 days	67.6	53.2
Jigsaw++ [4]	ImageNet	-	72.5	56.5
Shuffle and Learn	UCF-101	-	54.3	39.9
OPN [5]	UCF,HMDB,ACT	> 2 days	63.8	46.9
Our	UCF,HMDB,ACT	**12 hours**	**70.7**	**48.1**

[1] Doersch et al., “Unsupervised Visual Representation Learning by Context Prediction” (2015)

[2] Noroozi et al., “Representation Learning by Learning to Count” (2017)

[3] Noroozi et al., “Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles” (2016)

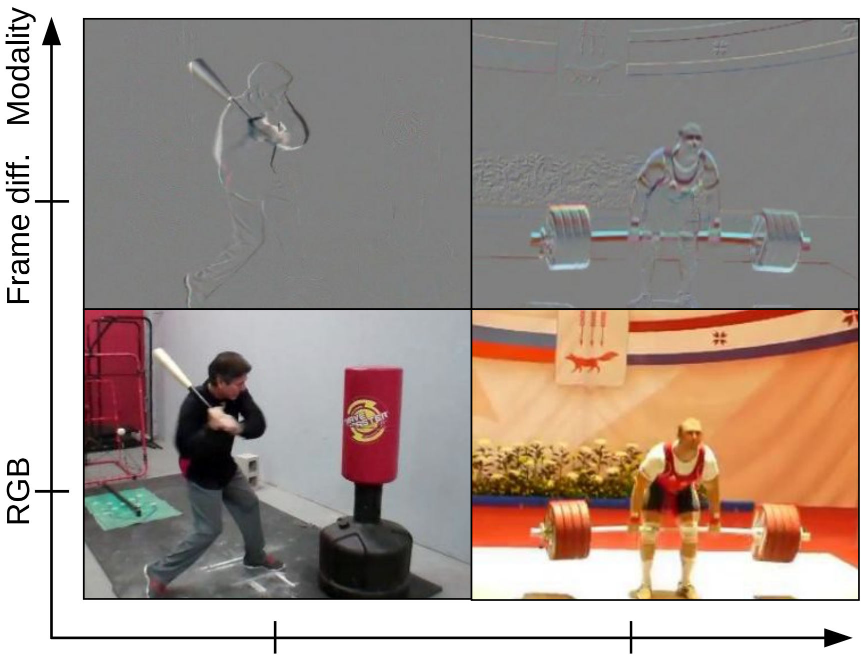
[4] Noroozi et al., “Boosting Self-Supervised Learning via Knowledge Transfer” (2018)

[5] Lee et al., “Unsupervised Representation Learning by Sorting Sequences” (2017)

Different Modalities

Frame differences as cheap alternative to optical flow

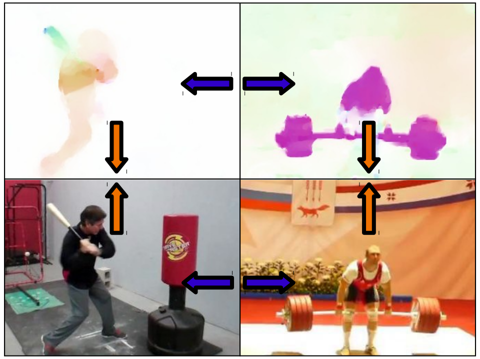
Benefit for all modalities

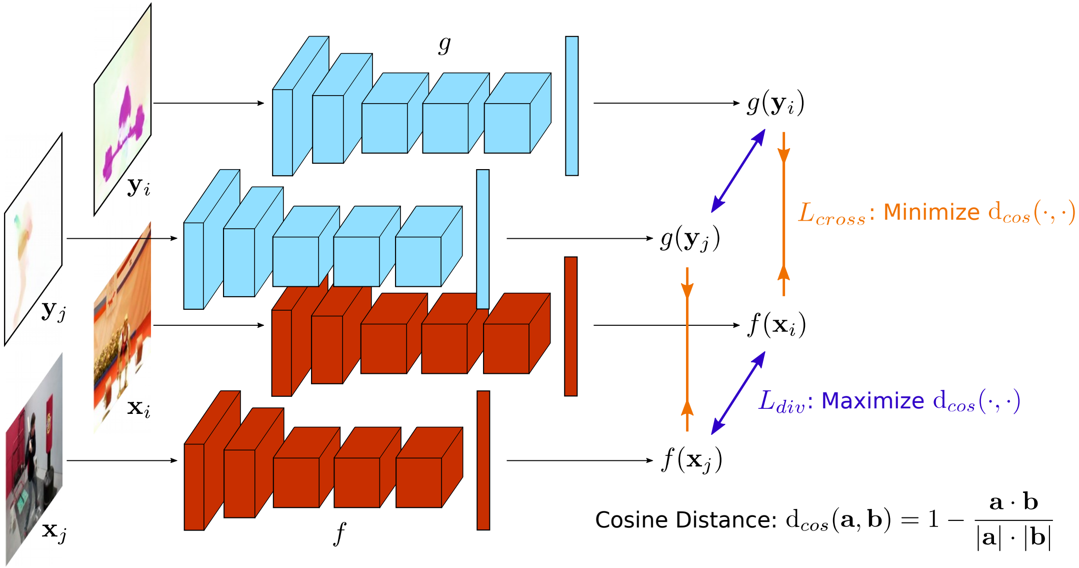


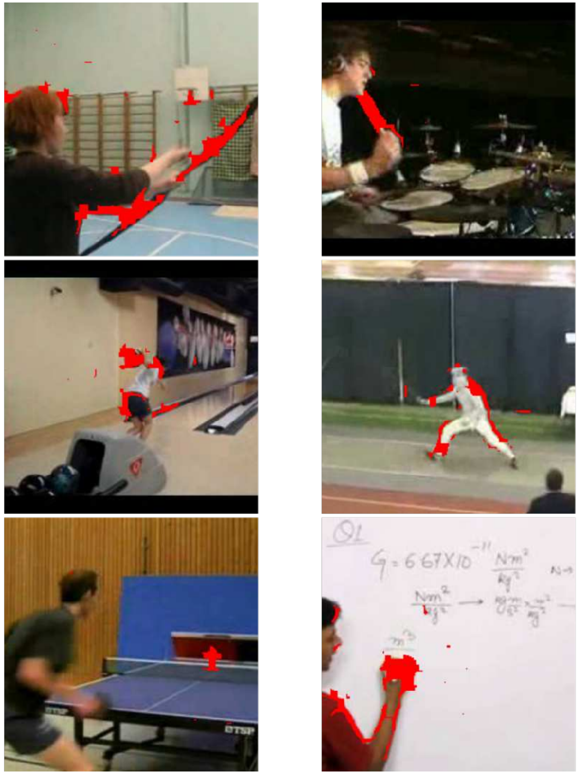
Action Recognition

Dataset	Pre-training	RGB &	Flow	RGB &	Frame diff.
UCF-101	No pre-training	49.1	76.4	49.1	64.5
UCF-101	Our pre-training	**59.3**	**79.2**	**55.4**	**66.3**
HMDB-51	No pre-training	19.2	47.1	19.2	30.8
HMDB-51	Our pre-training	**27.7**	**51.7**	**23.5**	**33.3**

Thank You! Questions?







Our model

https://hci.iwr.uni-heidelberg.de/compvis