Evolving Losses for Unsupervised Video Representation Learning

AJ Piergiovanni, Anelia Angelova, Michael S. Ryoo

Albert-Ludwigs-Universität Freiburg

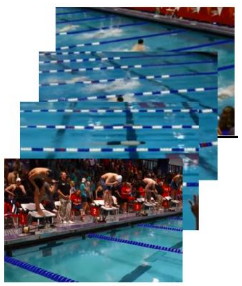
05.10.2020

Salem Ayedi

Advisor: David Hoffmann

Motivation

Videos



Smart Cities and homes

[1]

Robot Perception

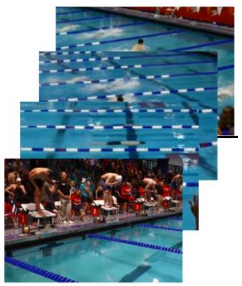
[2]

Web-Video Retrieval

[1] www.shutterstock.com, [2] https://twitter.com/mit/status/617031421958430721

Motivation

Videos



Smart Cities and homes

Robot Perception

Web-Video Retrieval

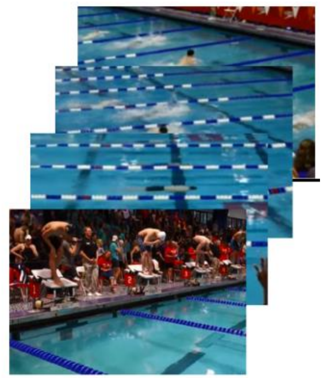
Learn from Videos ?

Motivation

Learn from Videos ?

Motivation

- Higher dimensions.

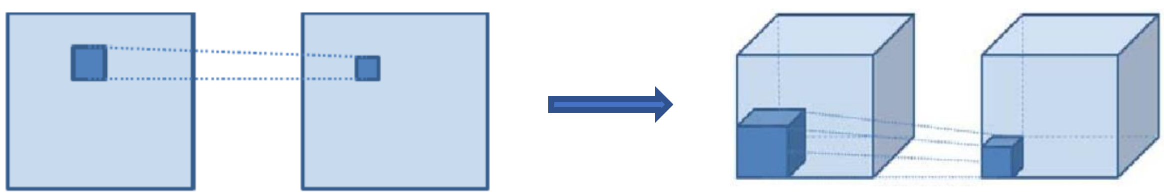


Learn from Videos ?

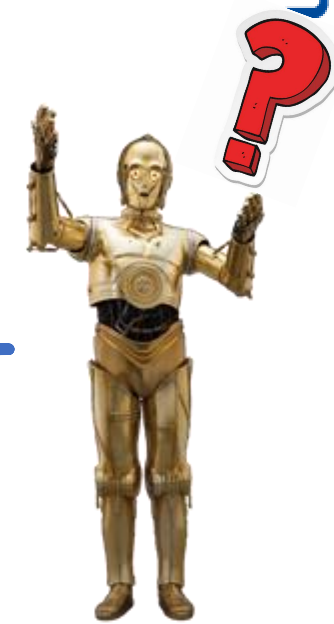
Motivation

- Higher dimensions.

- More trainable parameters if use 3D convs.



2D conv   3D conv



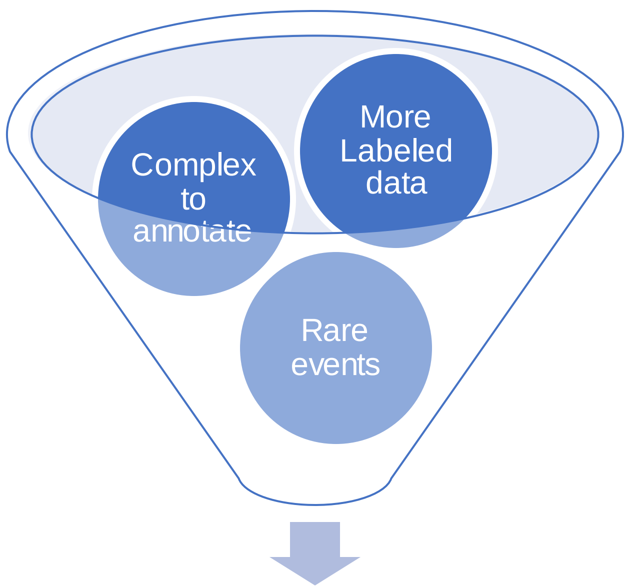
Learn from Videos ?

Motivation

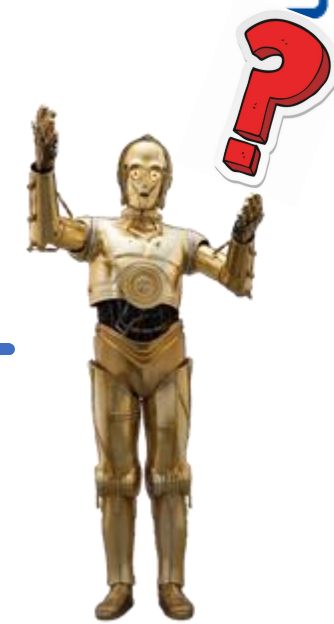
- Higher dimensions.

- More trainable parameters if use 3D convs.

- Expensive!



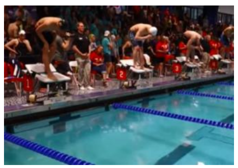
How to learn a good Video representation?

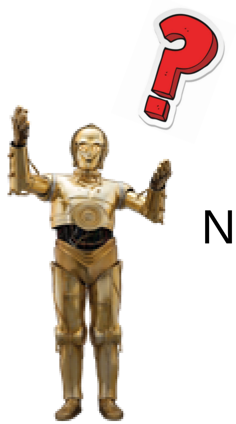


Learn from Videos ?

Goal

How to learn a robust video representation?





No labeled data

Not Domain Specific

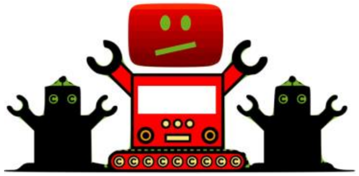
Generic

Transferrable

Goal

How to learn a robust video representation?

Unlabeled data



Unsupervised representation learning

Not Domain Specific

Generic

Transferrable

Goal

How to learn a robust video representation?

Unlabeled data

Unsupervised representation learning

Not Domain Specific

Generic

Transferrable

➔ Learn an **Unsupervised** representation by formulating an **Multi-Modal** and **Multi-task** learning problem.

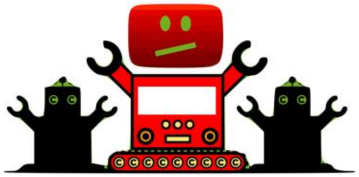
Video Modalities

Multiple Self-supervised tasks

Goal

How to learn a robust video representation?

Unlabeled data



Unsupervised representation learning

Not Domain Specific

Generic

Transferrable

➔ Learn an **Unsupervised** representation by formulating an **Multi-Modal** and **Multi-task** learning problem.

- Loss Function
- Evaluation Metric
- Single RGB Network

Plan

- Related work
- Approach:
  - Representation learning
  - Loss function
  - Evolving losses
  - Metrics
- Results

Plan

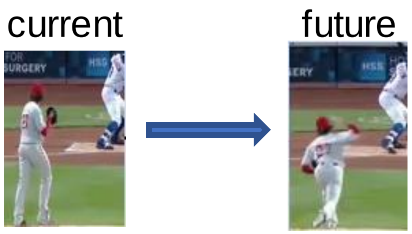
- Related work
- Approach:
  - Representation learning
  - Loss function
  - Evolving losses
  - Metrics
- Results

Related work

Self Supervised Learning for Video Representations:

Temporal structure

- Future prediction.
- Shuffled Frame Detection
- Forward/ Backward Detection

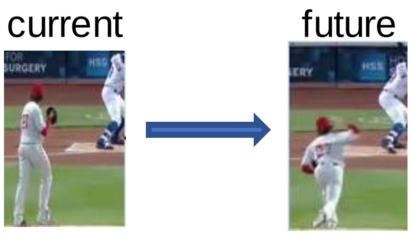


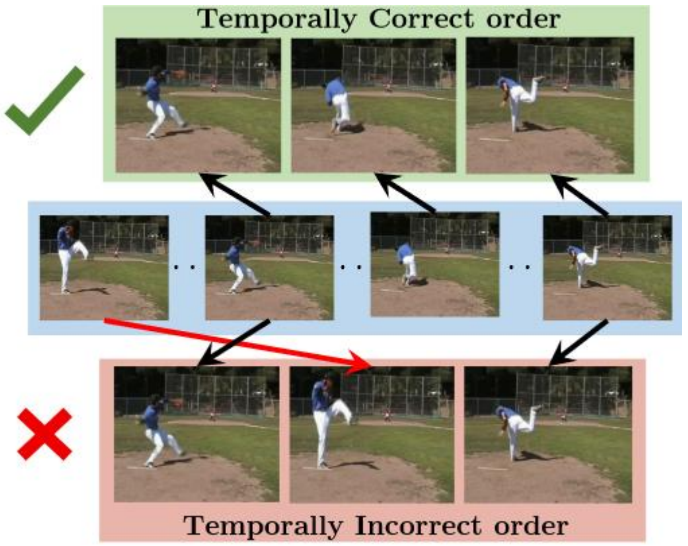
Related work

Self Supervised Learning for Video Representations:

Temporal structure

- Future prediction.
- Shuffled Frame Detection
- Forward/ Backward Detection





[1]

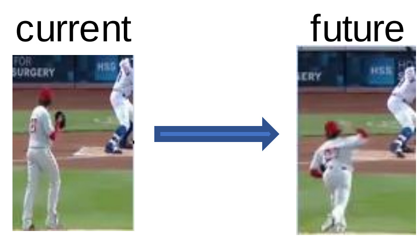
[1] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In Proceedings of European Conference on Computer Vision (ECCV), 2016.

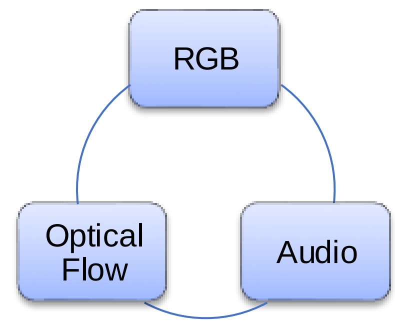
Related work

Self Supervised Learning for Video Representations:

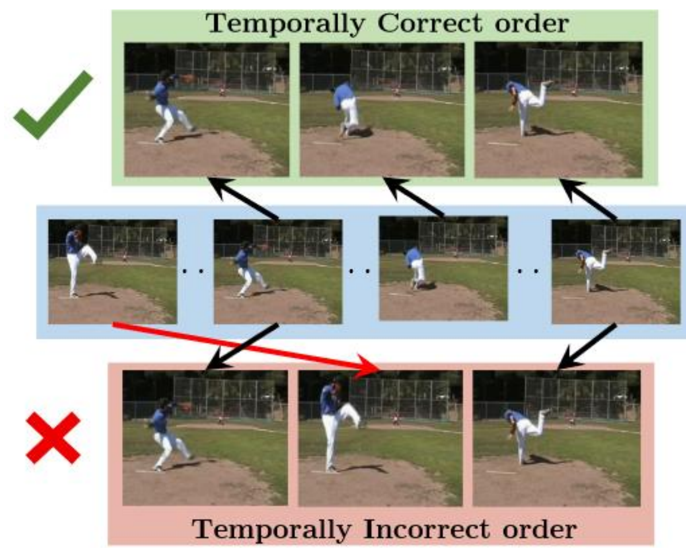
Temporal structure

- Future prediction.
- Shuffled Frame Detection
- Forward/ Backward Detection





[1]



[1] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In Proceedings of European Conference on Computer Vision (ECCV), 2016.

Related work

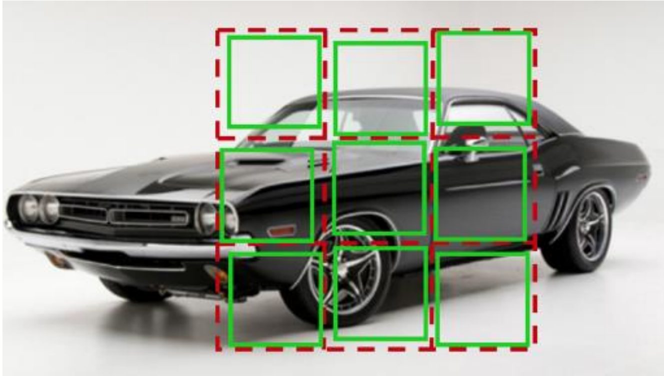
Self Supervised Learning for Video Representations:

Temporal structure

- Future prediction.
- Shuffled Frame Detection
- Forward/ Backward Detection

Spatial structure

- Tracking patches over time.
- Relative position patches detection



[2]

[2] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In Proceedings of European Conference on Computer Vision (ECCV), pages 69–84, 2016.

Related work

Self Supervised Learning for Video Representations:

Temporal structure

- Future prediction.
- Shuffled Frame Detection
- Forward/ Backward Detection

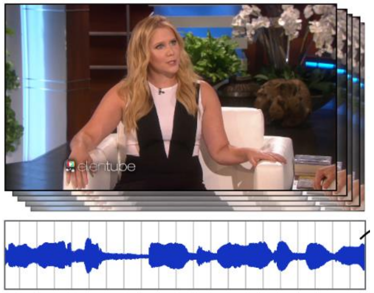
Spatial structure

- Tracking patches over time.
- Shuffled image parts

Multi Modal tasks

- Multi-Modal Alignment
- Cross Modal Translation

[3]



RGB to Flow

[3] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In Proceedings of European Conference on Computer Vision (ECCV), 2018..

Related work

Self Supervised Learning for Video Representations:

How do we learn a representation that combines all these tasks?

Related work

Multi-Task Self Supervised Learning:

- Future **RGB** prediction.
- Future **Audio** prediction.
- Shuffled **RGB** Detection
- Shuffled **Flow** Detection
- **Audio/RGB** Alignment
- **Flow/ RGB** Alignment

→ Representation

Carl Doersch and Andrew Zisserman. Multi-task selfsupervised visual learning. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.

Related work

Multi-Task Self Supervised Learning:

- Future **RGB** prediction.
- Future **Audio** prediction.
- Shuffled **RGB** Detection
- Shuffled **Flow** Detection
- **Audio/RGB** Alignment
- **Flow/ RGB** Alignment

→ Representation

! Tasks are assumed to have equal weights

➔ Learning from **multi-modal inputs** and **automatically** discovering the **weights** of the **tasks**

Carl Doersch and Andrew Zisserman. Multi-task selfsupervised visual learning. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.

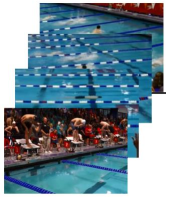
Plan

UNI FREIBURG

- Related work
- Approach:
  - Representation learning
  - Loss function
  - Evolving losses
  - Metrics
- Results

Approach: Overview

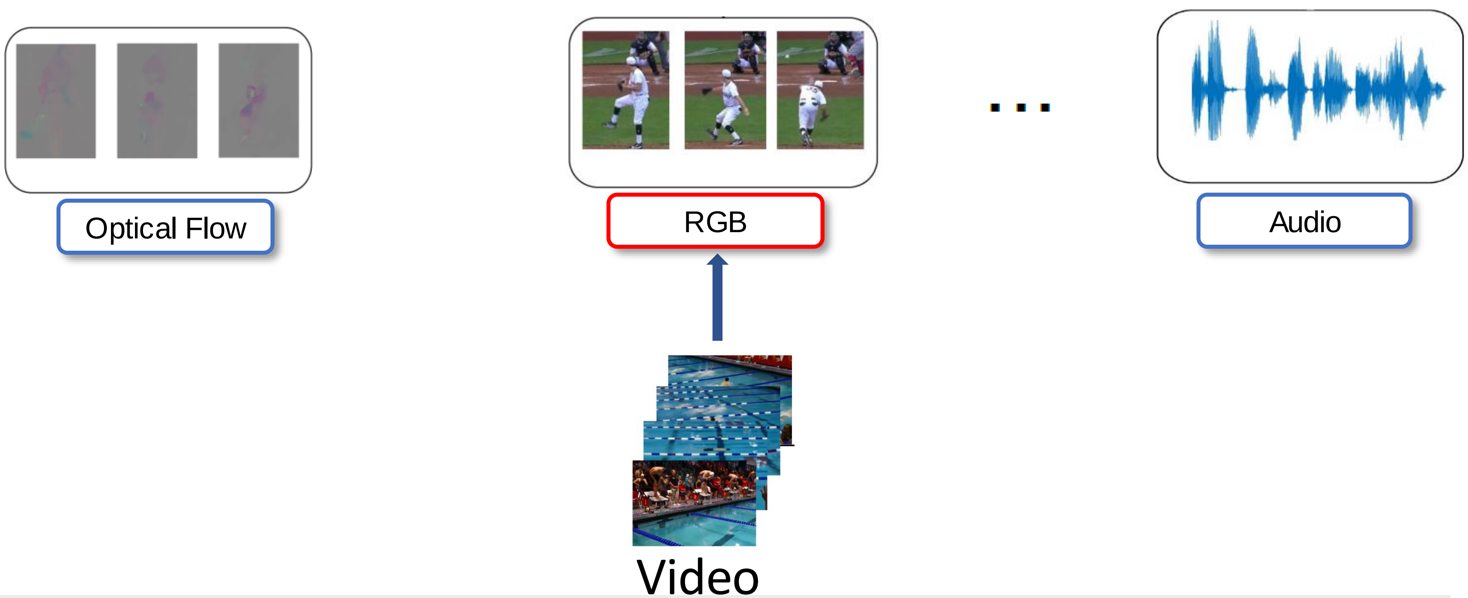
Input: unlabeled Video



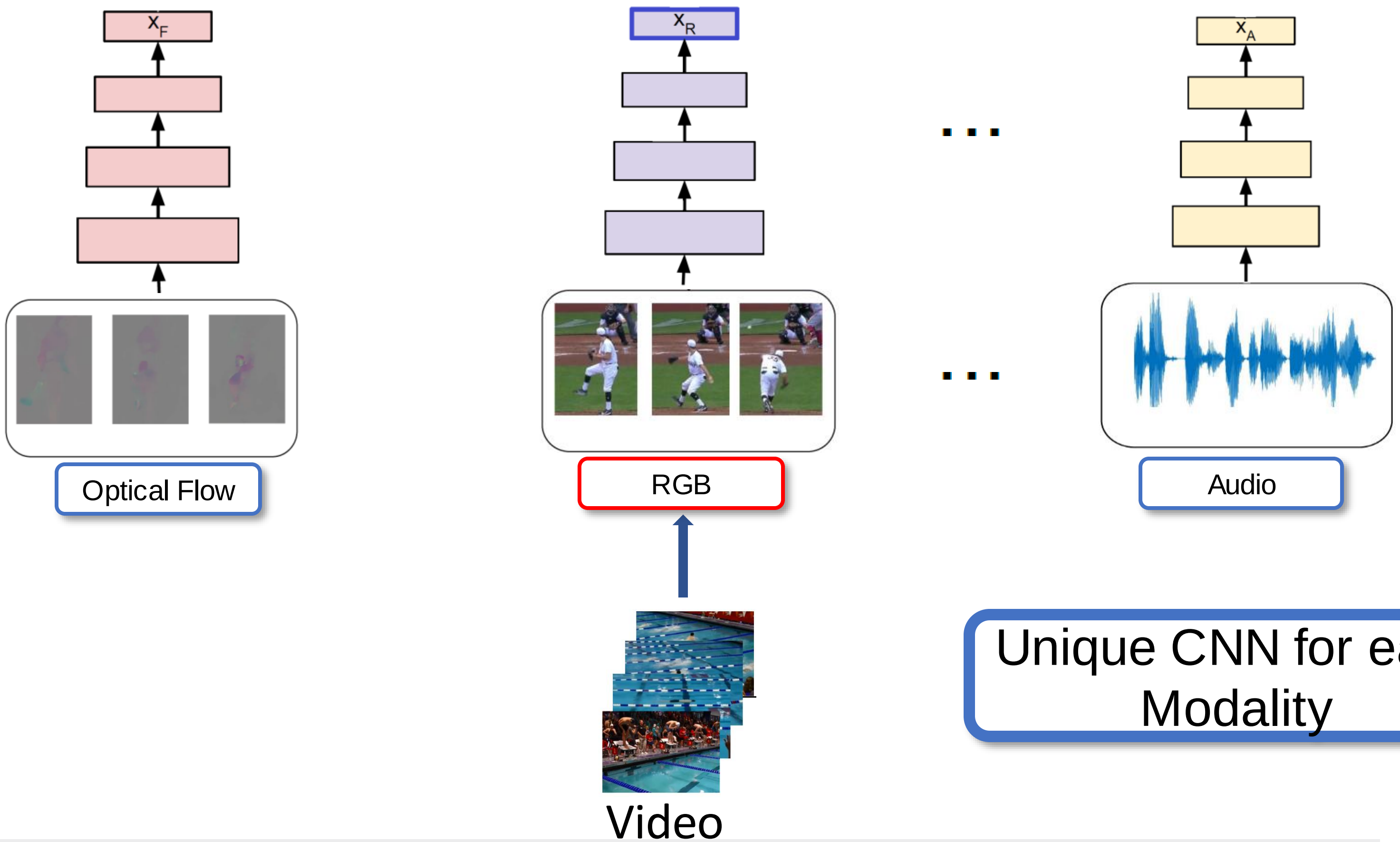
Video

Approach: Representation Learning

Modalities

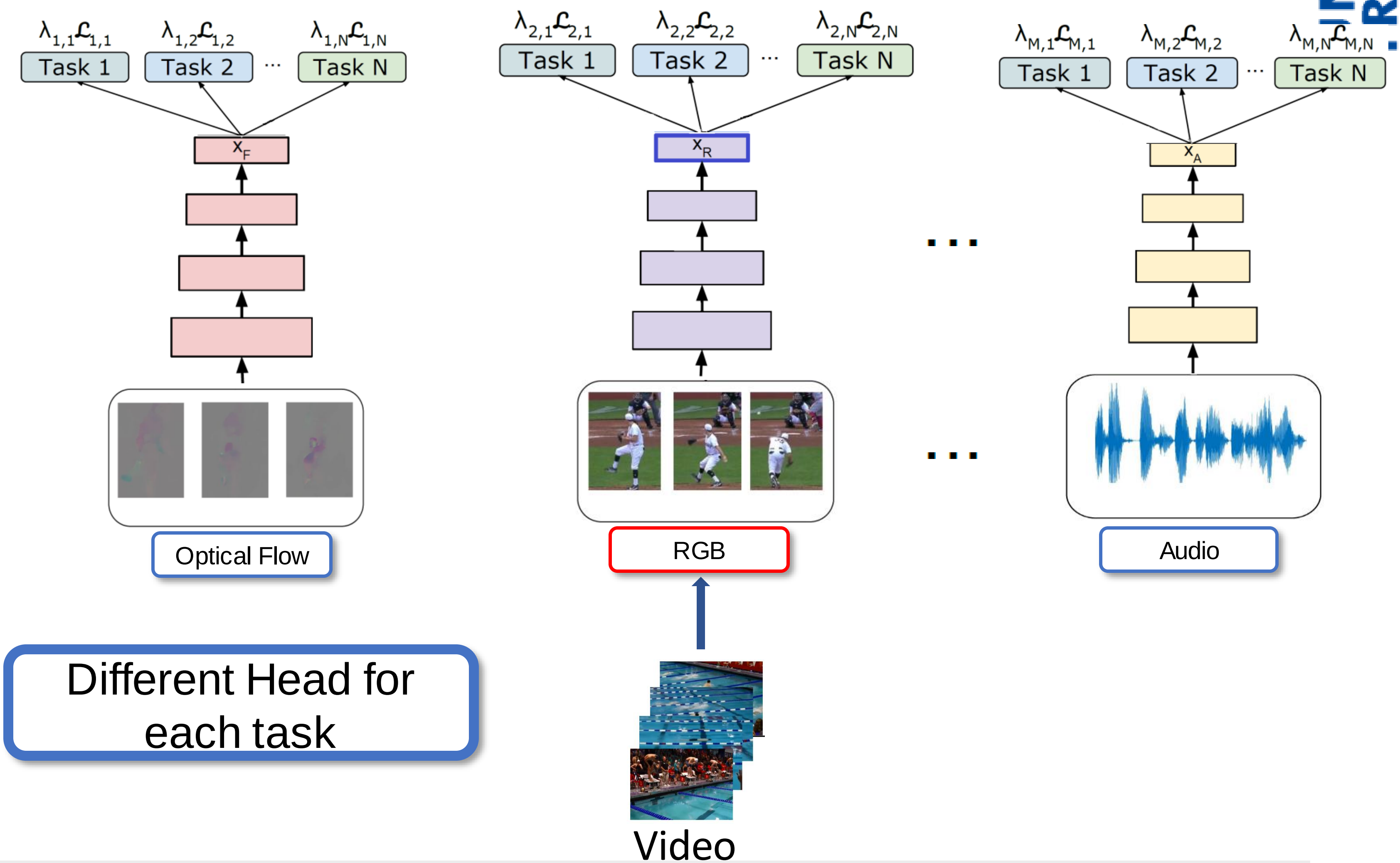


Approach: Representation Learning



Unique CNN for each Modality

Approach: Representation Learning



Different Head for each task

Approach: Representation Learning

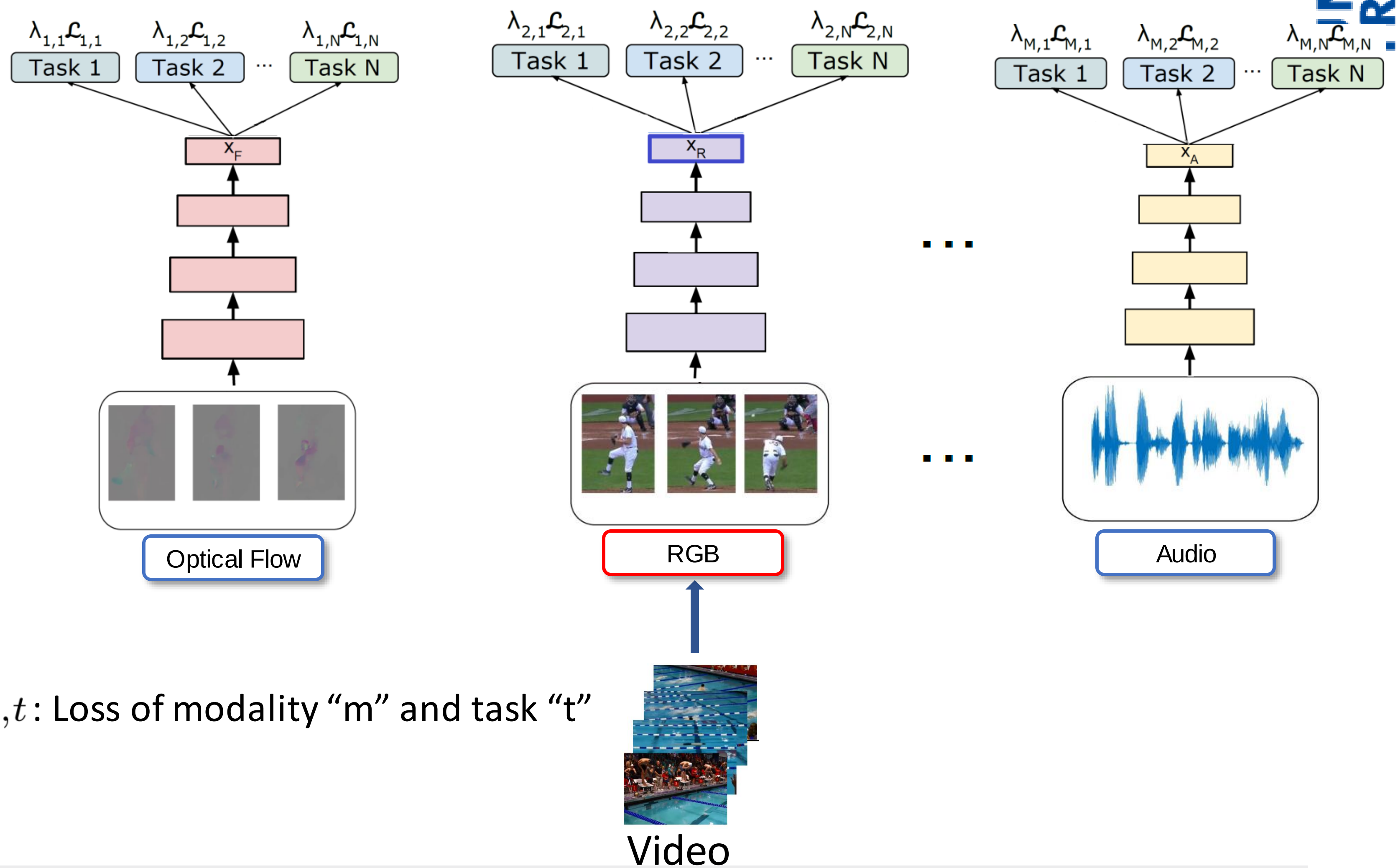
How to combine the information learned in each modality?

Approach: Representation Learning

How to combine the information learned in each modality?

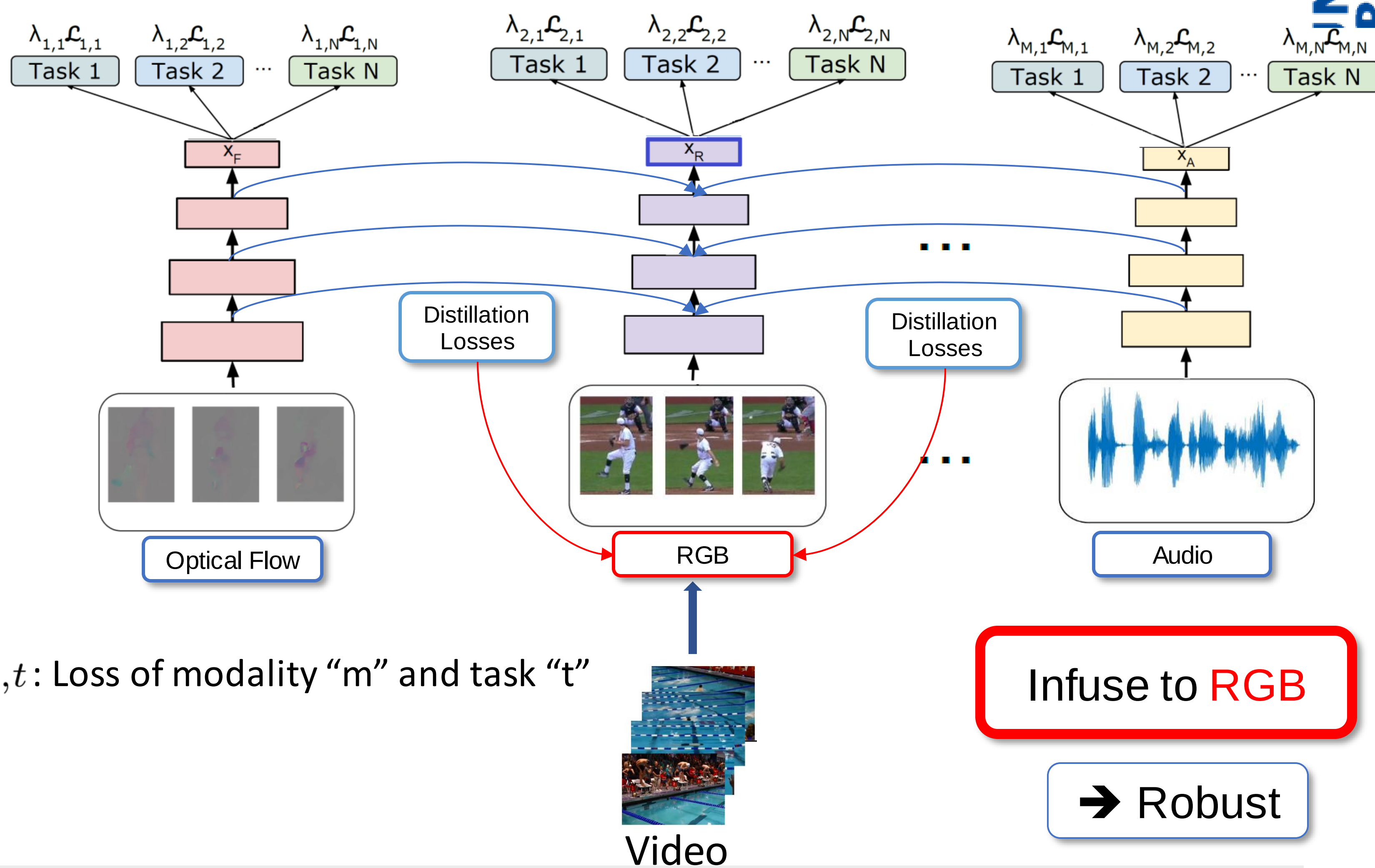
➔ “infuse” all the information to the RGB Network

Approach: Representation Learning



$\mathcal{L}_{m,t}$: Loss of modality “m” and task “t”

Approach: Representation Learning

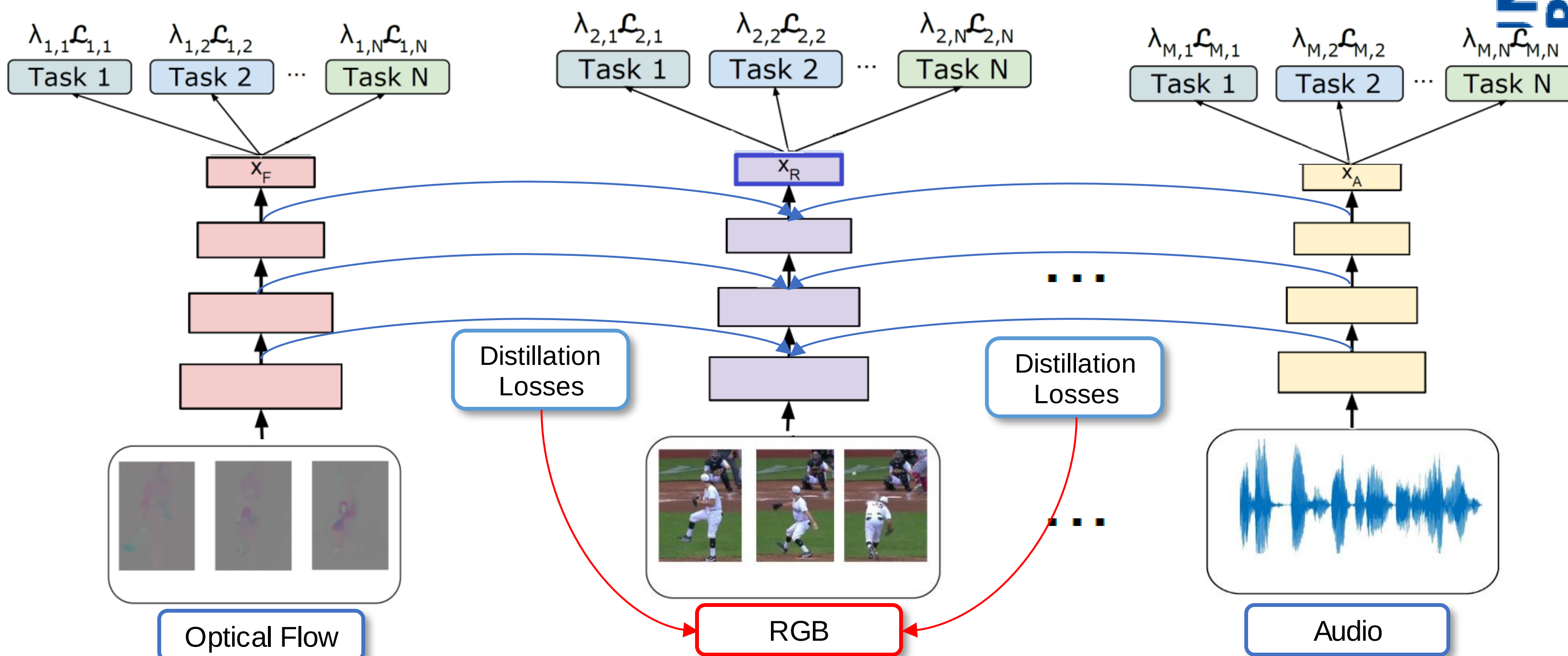


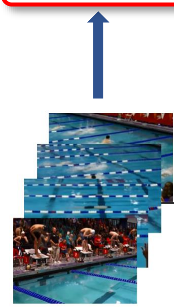
$\mathcal{L}_{m,t}$: Loss of modality “m” and task “t”

Infuse to RGB

➔ Robust

Approach: Representation Learning





Video

$\mathcal{L}_{m,t}$: Loss of modality "m" and task "t"

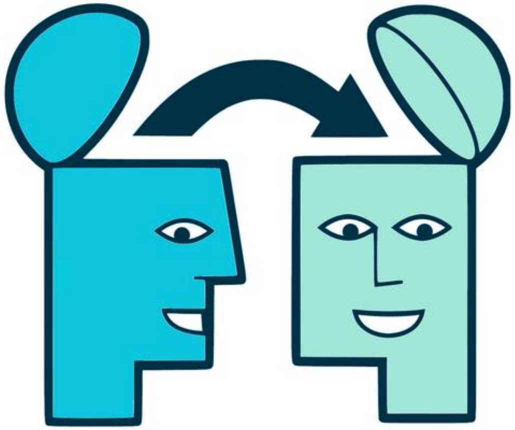
$\mathcal{L}_d$: Distillation Loss

Loss $\mathcal{L} = \sum_m \sum_t \lambda_{m,t}\mathcal{L}_{m,t} + \sum_d \lambda_d \mathcal{L}_d$

Approach: Loss function

$$\mathcal{L} = \sum_m \sum_t \lambda_{m,t} \mathcal{L}_{m,t} + \sum_d \lambda_d \mathcal{L}_d$$

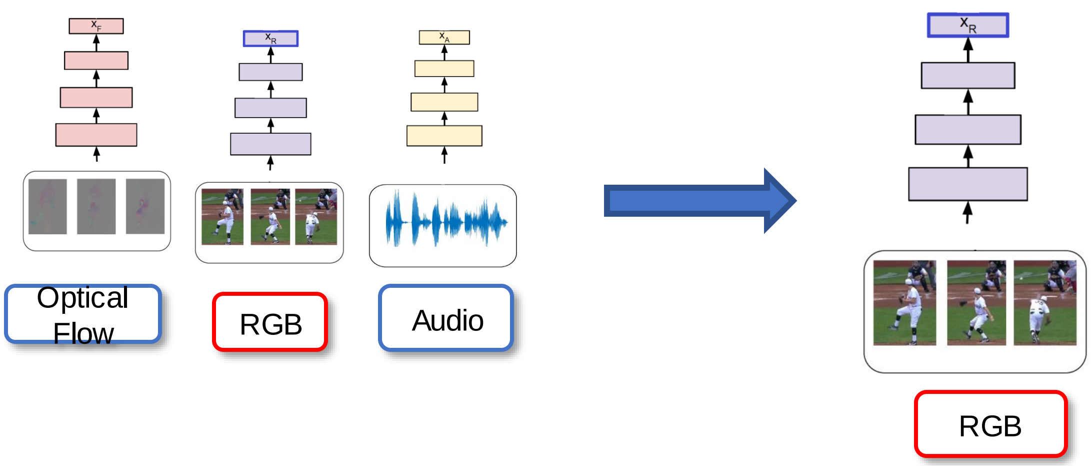
Distillation Loss $\mathcal{L}_d$ → Transfer Knowledge



Approach: Loss function

$$\mathcal{L} = \sum_{m} \sum_{t} \lambda_{m,t} \mathcal{L}_{m,t} + \sum_{d} \lambda_d \mathcal{L}_d$$

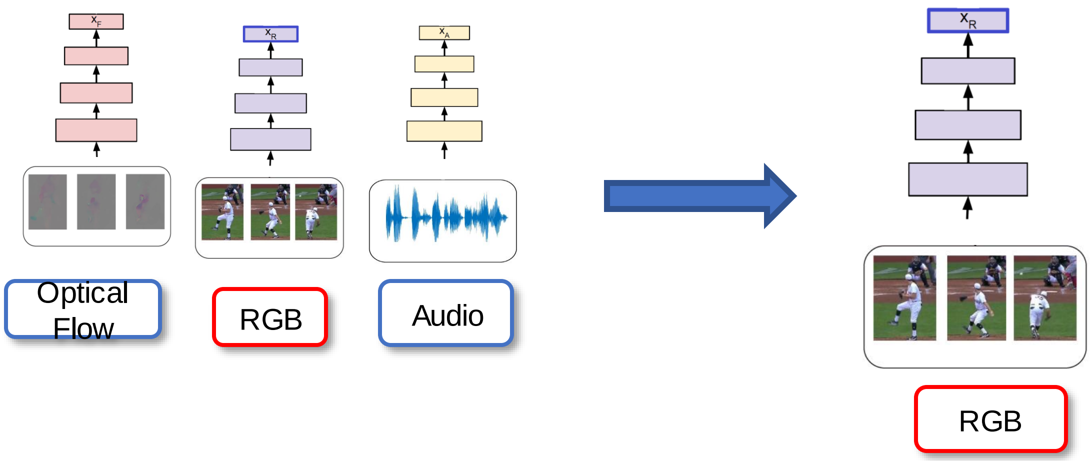
Distillation Loss $\mathcal{L}_d$ → Transfer Knowledge → Infuse



Approach: Loss function

$$\mathcal{L} = \sum_m \sum_t \lambda_{m,t} \mathcal{L}_{m,t} + \sum_d \lambda_d \mathcal{L}_d$$

Distillation Loss $\mathcal{L}_d$ → Transfer Knowledge → Infuse



$$\mathcal{L}_d(L_i, M_i) = ||L_i - M_i||_2$$

M_i : Activation of a layer in the **main** network

L_i : Activation of a layer of **another** network

Approach: Loss function

$$\mathcal{L} = \sum_{m} \sum_{t} \lambda_{m,t} \mathcal{L}_{m,t} + \sum_{d} \lambda_d \mathcal{L}_d$$

How to find these weights without any labeled data?

Approach: Evolving Loss function

$$\mathcal{L} = \sum_m \sum_t \lambda_{m,t} \mathcal{L}_{m,t} + \sum_d \lambda_d \mathcal{L}_d$$

Evolutionary Algorithms

Loss Population

$\lambda_{m,t}$ λ_d in $[0, 1]$

Approach: Evolving Loss function

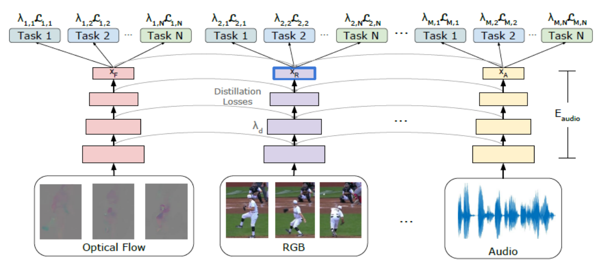
$$\mathcal{L} = \sum_m \sum_t \lambda_{m,t}\mathcal{L}_{m,t} + \sum_d \lambda_d \mathcal{L}_d$$

Evolutionary Algorithms

Loss Population

Train the networks of each Loss

$\lambda_{m,t}$ λ_d in $[0, 1]$



Approach: Evolving Loss function

$$\mathcal{L} = \sum_m \sum_t \lambda_{m,t} \mathcal{L}_{m,t} + \sum_d \lambda_d \mathcal{L}_d$$

Evolutionary Algorithms

Loss Population

Train the networks of each Loss

Evaluate each loss with the Fitness Criterion

$\lambda_{m,t}$ λ_d in $[0, 1]$

Approach: Evolving Loss function

$$\mathcal{L} = \sum_m \sum_t \lambda_{m,t} \mathcal{L}_{m,t} + \sum_d \lambda_d \mathcal{L}_d$$

Evolutionary Algorithms

Loss Population → Train the networks of each Loss → Evaluate each network with the Fitness Criteria → Mutate the top performing losses: Evolution Loss ➔ Child population

$\lambda_{m,t}$ λ_d in $[0, 1]$

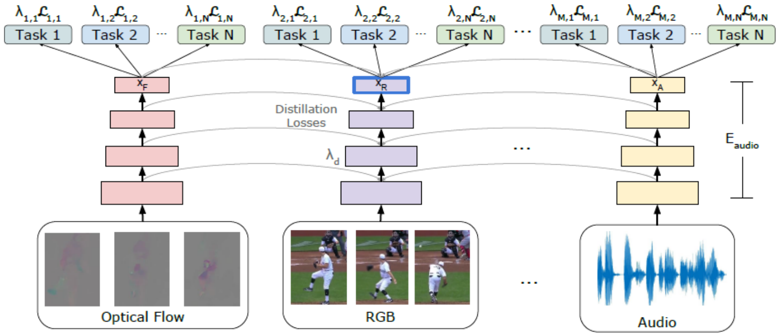
➔ Tournament Selection

➔ CMA-ES: Cov Matrix Adaptation

Approach: Summary ELo

1 – Define population of losses

2 – learn an unsupervised representation for each loss



3 – Evaluate how good is the learned representation of each loss

4 – Improve the loss generation

Approach: Evolving Loss function

$$\mathcal{L} = \sum_m \sum_t \lambda_{m,t} \mathcal{L}_{m,t} + \sum_d \lambda_d \mathcal{L}_d$$

Evolutionary Algorithms

Loss Population → Train the networks of each Loss → Evaluate each network with the Fitness Criteria → Mutate the top performing losses: Evolution Loss ➔ Child Population

$\lambda_{m,t}$ λ_d in $[0, 1]$

➔ Fitness Criteria

➔ Tournament Selection

➔ CMA-ES: Cov Matrix Adaptation

Approach: Evaluation Metric

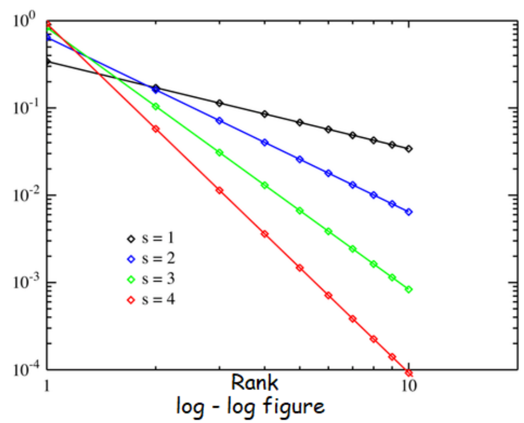
$$\mathcal{L} = \sum_m \sum_t \lambda_{m,t} \mathcal{L}_{m,t} + \sum_d \lambda_d \mathcal{L}_d$$

Fitness Criterion

→ Activity recognition

→ Zipf Distribution

$$q(c_i) = \frac{1/i^s}{H_{k,s}}$$



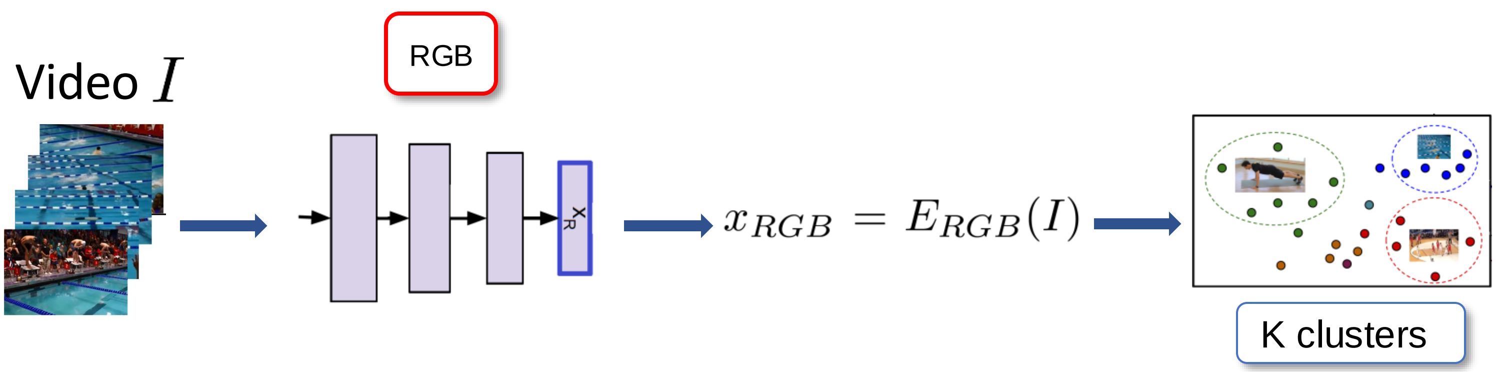
https://en.wikipedia.org/wiki/Zipf%27s_law

Approach: Evaluation Metric

$$\mathcal{L} = \sum_m \sum_t \lambda_{m,t} \mathcal{L}_{m,t} + \sum_d \lambda_d \mathcal{L}_d$$

Fitness Criterion

- ➔ Activity recognition
- ➔ Zipf Distribution



Approach: Evaluation Metric

$$\mathcal{L} = \sum_m \sum_t \lambda_{m,t} \mathcal{L}_{m,t} + \sum_d \lambda_d \mathcal{L}_d$$

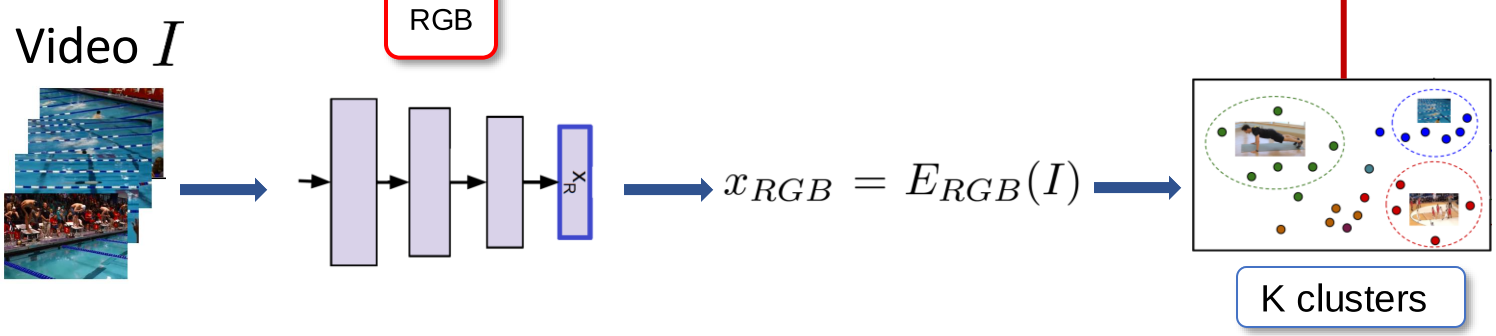
Fitness Criterion

➔ Fully Unsupervised

➔ Activity recognition

➔ Zipf Distribution

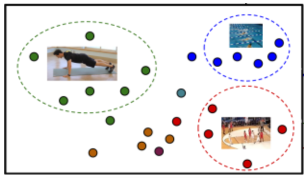
Compute KL Divergence



Approach: Evaluation Metric

$$\mathcal{L} = \sum_m \sum_t \lambda_{m,t} \mathcal{L}_{m,t} + \sum_d \lambda_d \mathcal{L}_d$$

KL Divergence



K clusters

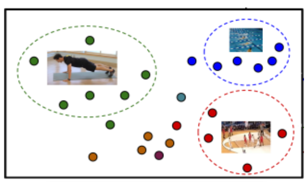
$$p(x|c_i) = \frac{1}{\sqrt{2\sigma^2\pi}} \exp\left(-\frac{(x-c_i)^2}{2\sigma^2}\right)$$

Likelihood of x in each class

Approach: Evaluation Metric

$$\mathcal{L} = \sum_m \sum_t \lambda_{m,t} \mathcal{L}_{m,t} + \sum_d \lambda_d \mathcal{L}_d$$

KL Divergence



K clusters

$$p(x|c_i) = \frac{1}{\sqrt{2\sigma^2\pi}} \exp\left(-\frac{(x-c_i)^2}{2\sigma^2}\right)$$

Likelihood of x in each class

$$p(c_i|x) = \frac{p(c_i)p(x|c_i)}{\sum_j^k p(c_j)p(x|c_j)}$$

$$= \frac{\exp-(x-c_i)^2}{\sum_{j=1}^k \exp|-(x-c_j)^2}$$

Equal prior for all clusters

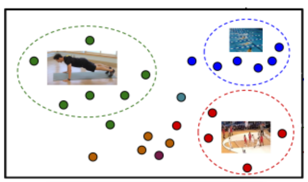
Bayes rule

$$p(c_i) = \frac{1}{N} \sum_{x \in V} p(c_i|x)$$

Approach: Evaluation Metric

$$\mathcal{L} = \sum_m \sum_t \lambda_{m,t} \mathcal{L}_{m,t} + \sum_d \lambda_d \mathcal{L}_d$$

KL Divergence



K clusters

$$p(x|c_i) = \frac{1}{\sqrt{2\sigma^2\pi}} \exp\left(-\frac{(x-c_i)^2}{2\sigma^2}\right)$$

Likelihood of x in each class

$$p(c_i|x) = \frac{p(c_i)p(x|c_i)}{\sum_j^k p(c_j)p(x|c_j)}$$

$$= \frac{\exp -(x-c_i)^2}{\sum_{j=1}^k \exp|-(x-c_j)^2}$$

Equal prior for all clusters → Bayes rule

$$p(c_i) = \frac{1}{N} \sum_{x \in V} p(c_i|x)$$

Compute **KL** Divergence

$$KL(p||q) = \sum_{i=1}^k p(c_i) \log\left(\frac{p(c_i)}{q(c_i)}\right)$$

Zipf Distribution $q(c_i)$

Approach: Evaluation Metric

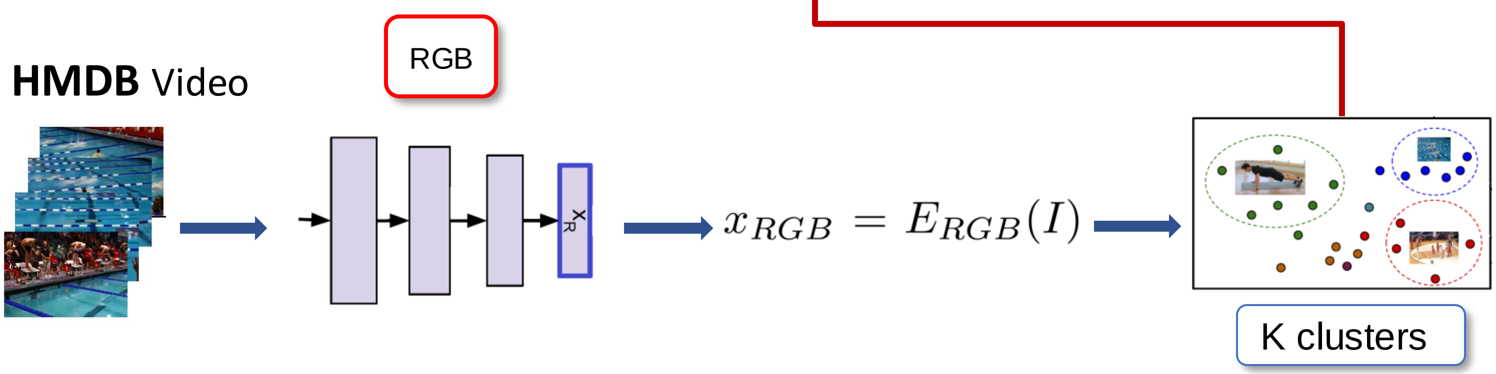
$$\mathcal{L} = \sum_{m} \sum_{t} \lambda_{m,t} \mathcal{L}_{m,t} + \sum_{d} \lambda_d \mathcal{L}_d$$

Fitness Criterion

➔Weakly Supervised

HMDB Labels

Accuracy



Plan

- Related work
- Approach:
    - Representation learning
    - Loss function
    - Evolving losses
    - Metrics
- Experiments and Results

Experiments and Results

UNI FREIBURG

Multi-Task Self Supervised Learning:

- Reconstruction tasks **for each modality**
- Future prediction **for each modality**.
- Temporal ordering **for each modality**.
- Cross-modality transfer tasks: **Flow to RGB…**
- **Multi-Modal** alignment
- **Multi-Modal** contrastive loss

Experiments and Results

Datasets

Training Dataset

2 Million **Random Unlabeled** Youtube Videos

Evaluation Dataset

HMDB, UCF101, Imagenet and Kinetics.

→ Less prune to bias and more general representation

Experiments and Results

Implementation Details

- (2+1)D ResNet 50 backbone network for each modality
- For a loss function, train the network for 100 epochs on 2 M videos
- During search, used smaller networks (ResNet-18); the fitness of each model can be found in 4 hours using 8 GPUs
- The final model uses 64 GPUs for 3 days

Experiments and Results

Method	HMDB	UCF101
Supervised		
(2+1)D ResNet-50 Scratch	35.2	63.1
(2+1)D ResNet-50 ImageNet	49.8	84.5
(2+1)D ResNet-50 Kinetics	74.3	95.1
Unsupervised		
Shuffle [26]	18.1	50.2
O3N [12]	32.5	60.3
OPN [24]	37.5	37.5
Patch [43]	-	41.5
Multisensory [29]	-	82.1
AVTS [22]	61.6	89.0
Weakly guided, HMDB		
Evolved Loss (ours)	67.8	94.1
Unsupervised		
Evolved Loss (ours, no distiliation)	53.7	84.2
Evolved Loss - ELo (ours)	67.4	93.8

Table 2: Comparison to SoTA on HMDB51 and UCF101

➔Importance of distillation

Experiments and Results

Method	HMDB	UCF101
Supervised		
(2+1)D ResNet-50 Scratch	35.2	63.1
(2+1)D ResNet-50 ImageNet	49.8	84.5
(2+1)D ResNet-50 Kinetics	74.3	95.1
Unsupervised		
Shuffle [26]	18.1	50.2
O3N [12]	32.5	60.3
OPN [24]	37.5	37.5
Patch [43]	-	41.5
Multisensory [29]	-	82.1
AVTS [22]	61.6	89.0
Weakly guided, HMDB		
Evolved Loss (ours)	67.8	94.1
Unsupervised		
Evolved Loss (ours, no distiliation)	53.7	84.2
Evolved Loss - ELo (ours)	67.4	93.8

Table 2: Comparison to SoTA on HMDB51 and UCF101

Experiments and Results

Method	HMDB	UCF101
Supervised		
(2+1)D ResNet-50 Scratch	35.2	63.1
(2+1)D ResNet-50 ImageNet	49.8	84.5
(2+1)D ResNet-50 Kinetics	74.3	95.1
Unsupervised		
Shuffle [26]	18.1	50.2
O3N [12]	32.5	60.3
OPN [24]	37.5	37.5
Patch [43]	-	41.5
Multisensory [29]	-	82.1
AVTS [22]	61.6	89.0
Weakly guided, HMDB		
Evolved Loss (ours)	67.8	94.1
Unsupervised		
Evolved Loss (ours, no distiliation)	53.7	84.2
Evolved Loss - ELo (ours)	67.4	93.8

Table 2: Comparison to SoTA on HMDB51 and UCF101

Experiments and Results

Method	k-means	1-layer	fine-tune
Supervised using additional labeled data			
Scratch (No Pretraining)	15.7	17.8	35.2
ImageNet Pretrained	32.5	37.8	49.8
Kinetics Pretrained	68.8	71.5	74.3
Unsupervised using unlabeled videos			
Frame Shuffle [26]	22.3	24.3	28.4
Reverse Detection [31]	21.3	24.3	27.5
Audio/RGB Align [29, 22]	32.4	36.8	40.2
RGB to Flow	31.5	36.4	39.9
Predicting 4 future frames	31.8	35.8	39.2
Joint Embedding	29.4	32.5	38.4
Ours, weakly-sup clustering, using unlabeled videos			
Evolved Loss - ELo-weak	45.7	64.3	67.8
Ours, unsupervised, using unlabeled videos			
Random Loss (unsup.)	26.4	26.9	31.2
Evolved Loss - ELo (unsup.)	43.4	64.5	67.4

Table 1: Evaluation of various self-supervised methods on HMDB51

Experiments and Results

Method	k-means	1-layer	fine-tune
Supervised using additional labeled data			
Scratch (No Pretraining)	15.7	17.8	35.2
ImageNet Pretrained	32.5	37.8	49.8
Kinetics Pretrained	68.8	71.5	74.3
Unsupervised using unlabeled videos			
Frame Shuffle [26]	22.3	24.3	28.4
Reverse Detection [31]	21.3	24.3	27.5
Audio/RGB Align [29, 22]	32.4	36.8	40.2
RGB to Flow	31.5	36.4	39.9
Predicting 4 future frames	31.8	35.8	39.2
Joint Embedding	29.4	32.5	38.4
Ours, weakly-sup clustering, using unlabeled videos			
Evolved Loss - ELo-weak	45.7	64.3	67.8
Ours, unsupervised, using unlabeled videos			
Random Loss (unsup.)	26.4	26.9	31.2
Evolved Loss - ELo (unsup.)	43.4	64.5	67.4

Table 1: Evaluation of various self-supervised methods on HMDB51

→ Importance of Evolution Loss

Experiments and Results

$$\mathcal{L} = \sum_m \sum_t \lambda_{m,t} \mathcal{L}_{m,t} + \sum_d \lambda_d \mathcal{L}_d$$

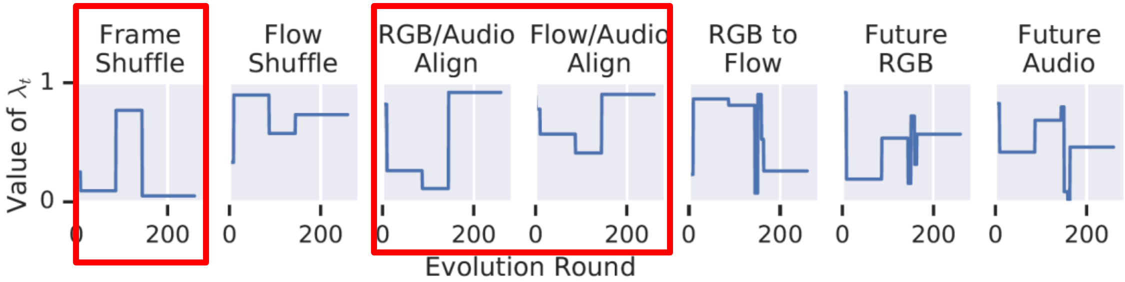
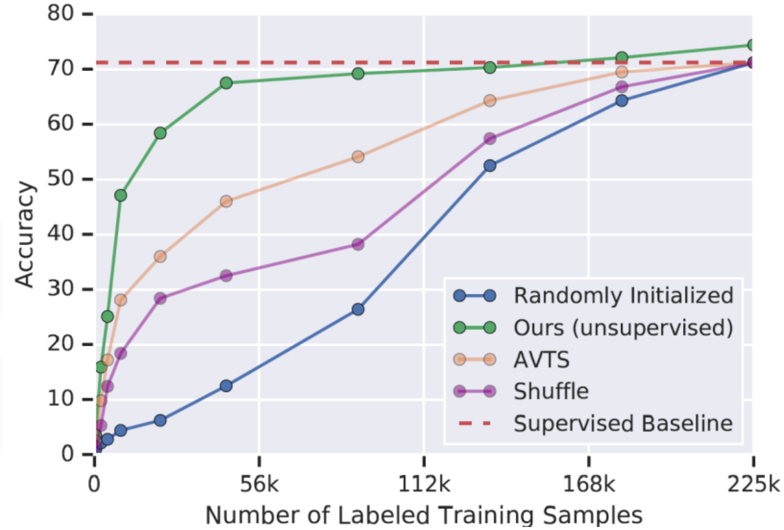


Figure 7: The values of the loss function for the various tasks throughout evolution

Experiments and Results

➔Improving Supervised Learning

Because you start with a good representation



Method	400	2k	4k	8k	20k	40k	80k	120k	160k	225k (all samples)
	Number of Labeled Samples									
Random Init	0.93	2.1	2.8	4.4	6.2	12.5	26.4	52.5	64.3	71.2
Frame Shuffle	1.5	5.3	12.4	18.4	28.4	32.5	38.2	57.4	66.8	70.9
Audio Align	2.5	9.8	17.2	28.1	36.0	46.0	54.1	64.3	69.5	71.5
ELo (unsupervised)	3.6	15.8	24.8	47.0	58.3	67.5	69.2	**70.2**	**72.2**	**74.4**

Figure 5 and Table 3: How much Labeled, supervised data to achieve SoTA

Experiments and Results

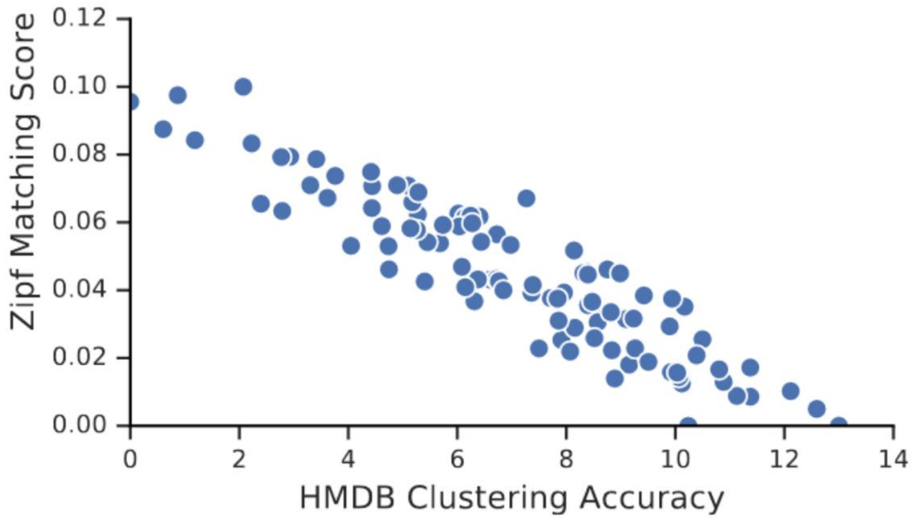


Figure 9: Comparison of the fitness measures for 100 different loss functions

➔Strong Correlation

➔Zipf matching is suitable for unsupervised representation evaluation

Experiments and Results

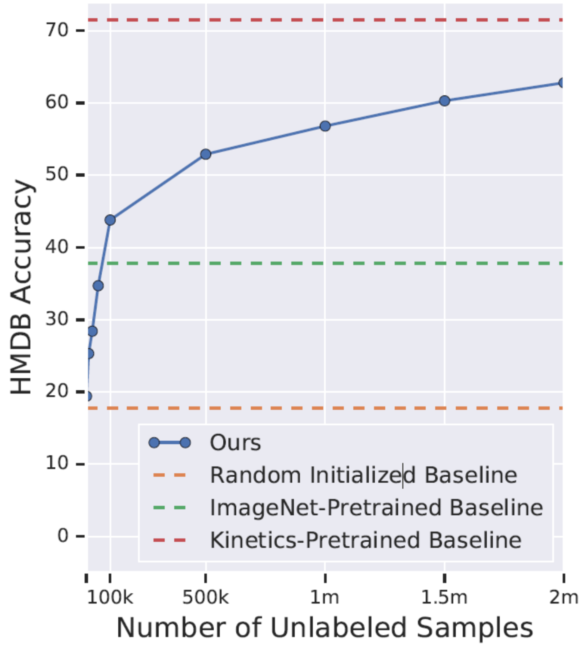
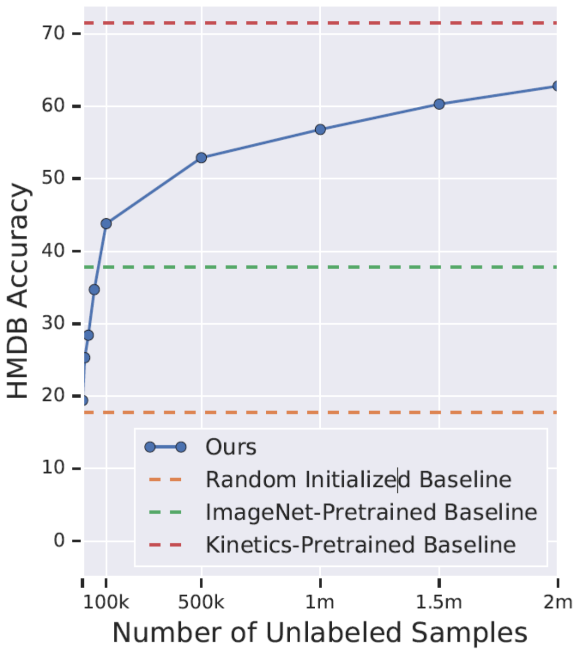


Figure 6: Different amounts of unsupervised data

Experiments and Results



Method	HMDB	UCF101
Supervised		
(2+1)D ResNet-50 Scratch	35.2	63.1
(2+1)D ResNet-50 ImageNet	49.8	84.5
(2+1)D ResNet-50 Kinetics	74.3	95.1
Unsupervised		
Shuffle [26]	18.1	50.2
O3N [12]	32.5	60.3
OPN [24]	37.5	37.5
Patch [43]	-	41.5
Multisensory [29]		82.1
AVTS [22]	$\boxed{61.6}$	89.0
Weakly guided, HMDB		
Evolved Loss (ours)	67.8	94.1
Unsupervised		
Evolved Loss (ours, no distiliation)	53.7	84.2
Evolved Loss - ELo (ours)	67.4	93.8

Figure 6: Different amounts of unsupervised data

Conclusion

- Formulate an **unsupervised** video representation as **Multi-Modal** and **Multi-task** learning problem.
- **Infuse** the information to RGB network
- **loss** function **evolution**
- **unsupervised fitness**

➔ Powerful video representation.

➔ Match or improve the performance of networks trained on supervised data

Thank you !

Bibliography

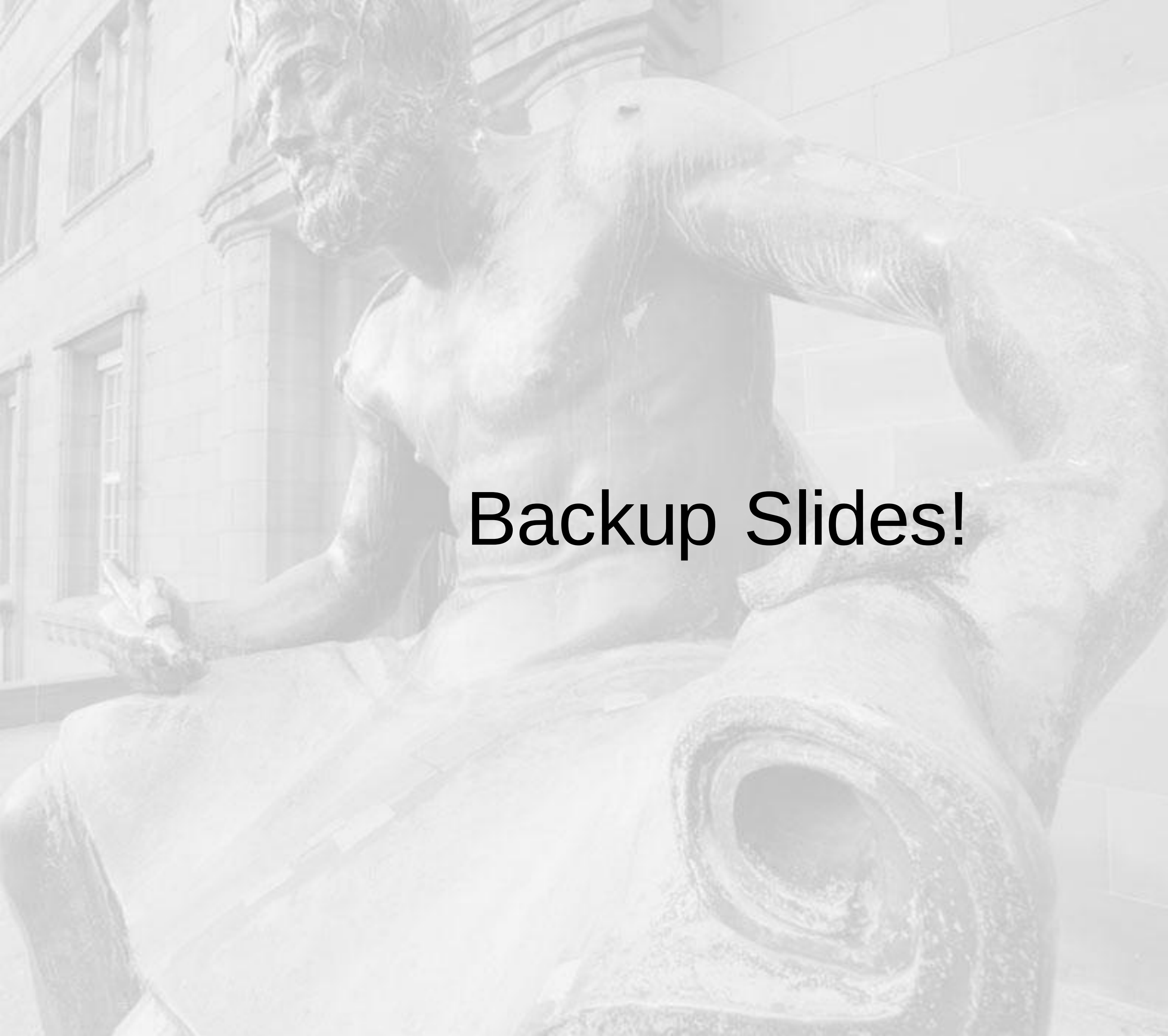
AJ Piergiovanni, Anelia Angelova, Michael S. Ryoo: *Evolving Losses for Unsupervised Video Representation Learning*

Ishan Misra, C Lawrence Zitnick, and Martial Hebert. *Shuffle and learn: unsupervised learning using temporal order verification. In Proceedings of European Conference on Computer Vision (ECCV), 2016.*

Mehdi Noroozi and Paolo Favaro. *Unsupervised learning of visual representations by solving jigsaw puzzles. In Proceedings of European Conference on Computer Vision (ECCV), pages 69–84, 2016.*

Andrew Owens and Alexei A Efros. *Audio-visual scene analysis with self-supervised multisensory features. In Proceedings of European Conference on Computer Vision (ECCV), 2018..*

Carl Doersch and Andrew Zisserman. *Multi-task selfsupervised visual learning. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.*



Backup Slides!

CMA ES

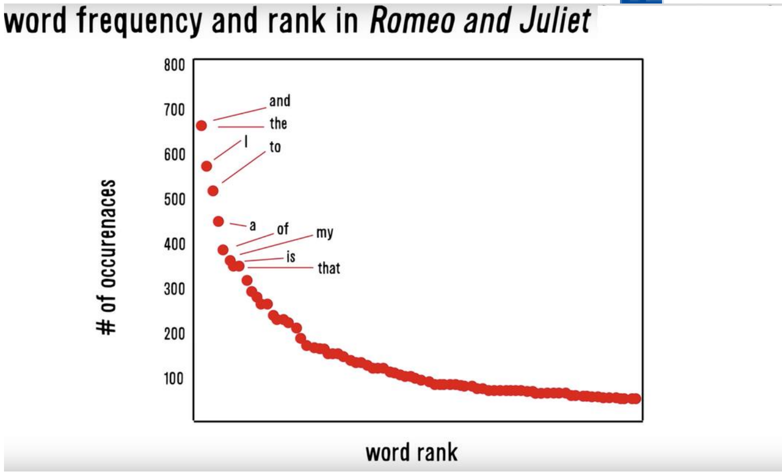
- Offspring not generated by the mutation of each single individual:
  - Choose random j: $x_i = x_j + \lambda_i z$
- But from weighted mean of the current population
  - $x_i = mean + \lambda_i z$
- With z ~ $\mathcal{N}(0, C)$ and C is the covariance matrix

Zipf distribution

$$q(c_i) = \frac{1/i^s}{H_{k,s}}$$

- Generalized Harmoic number

$$H_{k,s} = \frac{1/k^s}{\sum_{n=1}^{N}(1/n^s)}$$



Where:

- „N“: number of elements
- „i“: is the rank

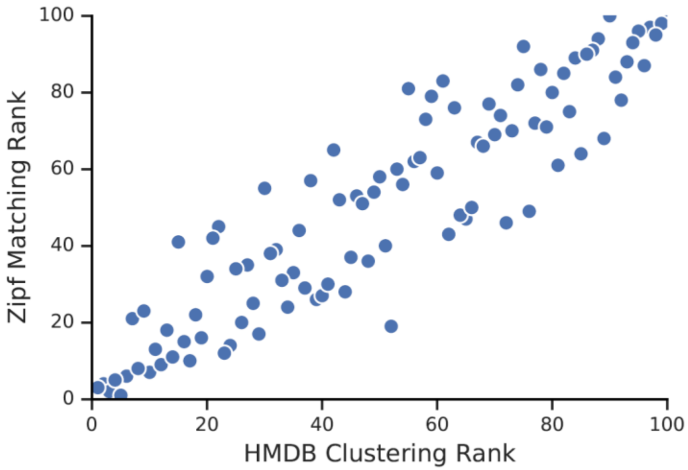


Figure 9: Comparison of the fitness measures for 100 different loss functions