ROBOVIS 2024

4th International Conference on Robotics, Computer Vision and Intelligent Systems

Rome, Italy 25 - 27 February, 2024

Learning with Privileged Information and Distillation for Multimodal Video Classification

Vittorio Murino

UNIVERSITÀ di VERONA
Dipartimento di INFORMATICA

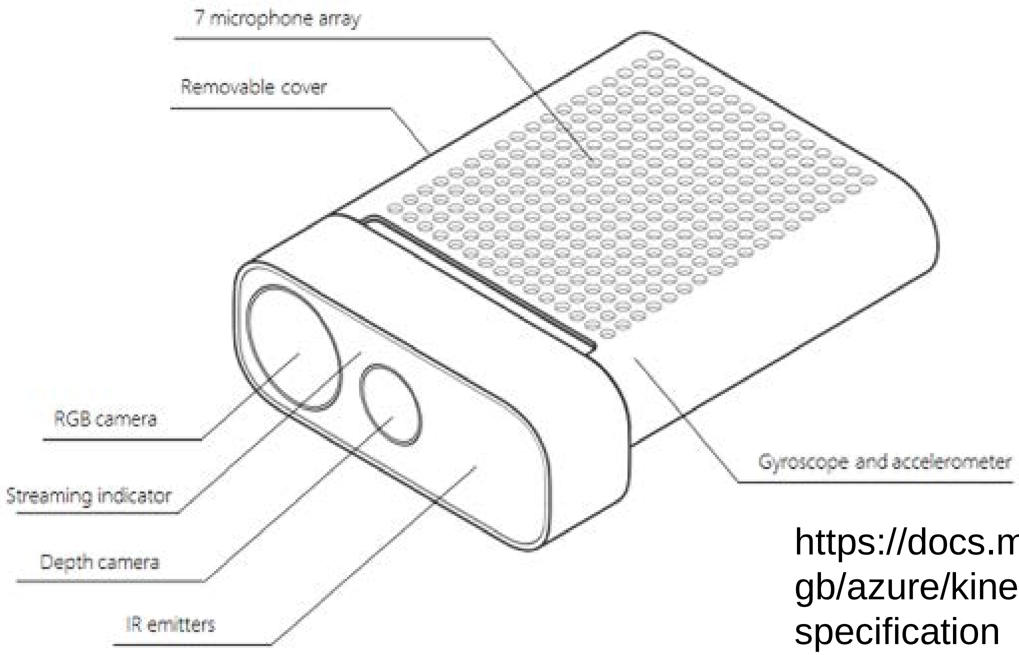
Dibris
UNIVERSITÀ DEGLI STUDI DI GENOVA

iit PAVIS
ISTITUTO ITALIANO DI TECNOLOGIA
PATTERN ANALYSIS AND COMPUTER VISION

Rome, February 27th, 2024

Multimodal Data

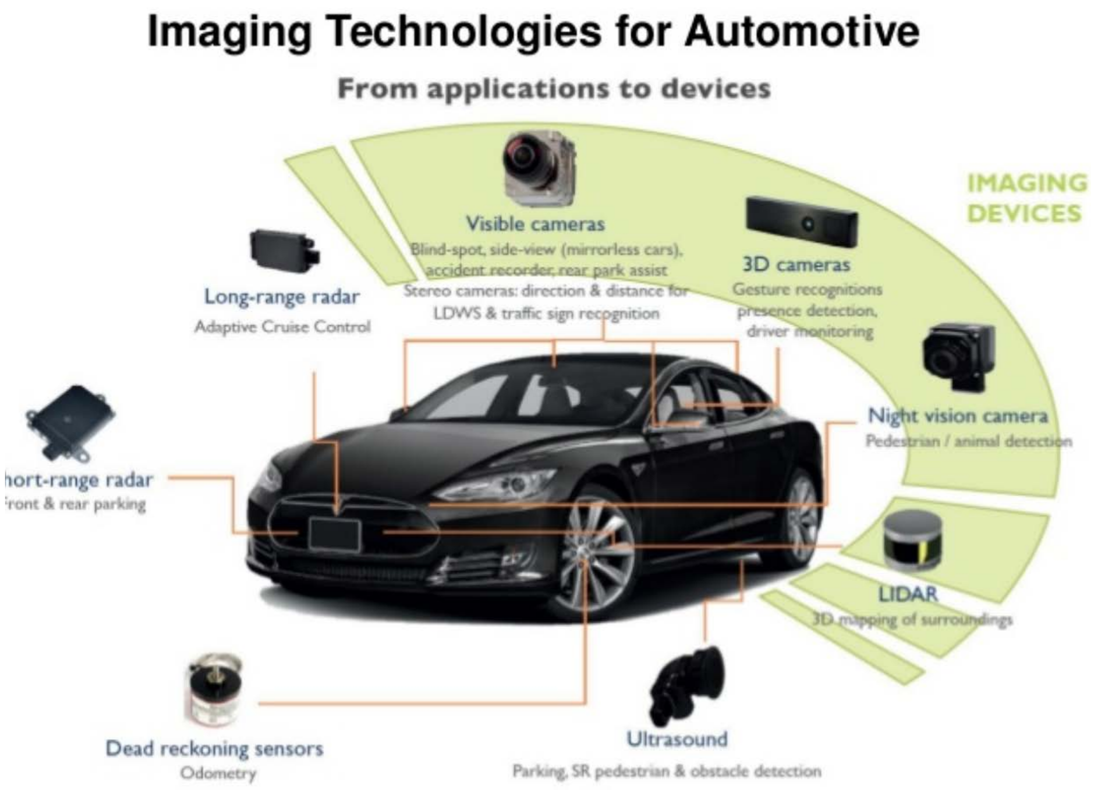
Multimodal vs. multisensory



https://docs.microsoft.com/en-gb/azure/kinect-dk/hardware-specification

Liu, Jun, et al. "NTU RGB+ D 120: A Large-Scale Benchmark for 3D Human Activity Understanding." IEEE TPAMI (2019)

Multimodal Data



Multimodal Learning

- More data
- More variety of data
- More (semantic) information (e.g., optical flow, joints, etc.)

⇨ Multiple modalities bring complementary information

But more data to process !

Outline

- Multimodal learning
- Privileged Information and Distillation
- A possible strategy, HALLUCINATION networks: two approaches
- Wrap-up and take-home message

Multimodal Learning

Challenges of Multimodal Learning:

1. How can we build deep learning models that learn on these different clues.
    - Issue: balancing the learning pace of the modalities

2. How can these deep learning models be used in case of a missing modality?

Multimodal Learning

- Assuming that we have RGB only available at test time, we can cast this question in another way:

  - How can other modalities help in learning a better RGB model?

⇨ Distillation + Privileged Information framework ⇦

Missing Modality in multimodal Video Action Recognition

Learning with Privileged Information

1. Train a model exploiting <u>multimodal data</u>

TRAINING

Depth

RGB

→ Handshaking

2. How to deal with a <u>missing modality</u> at test time?

TESTING

... ...

RGB

→ Handshaking

Liu, Jun, et al. "NTU RGB+ D 120: A Large-Scale Benchmark for 3D Human Activity Understanding." IEEE TPAMI (2019)

Main strategy

- Improve single-modality system performance using side information: *Privileged Information and distillation* → *generalized distillation*

- Use this extra modality in training only to *extract* suitable information

- Strategy: *Hallucination* networks
  - Trying to mimic the missing modality
  - Not necessarily at level of data, but at feature and prediction levels
  - *Distill* useful info from the missing modality data stream

Lopez-Paz, D., Bottou, L., Schölkopf, B., Vapnik, V. *Unifying distillation and privileged information*. ICLR 2016.

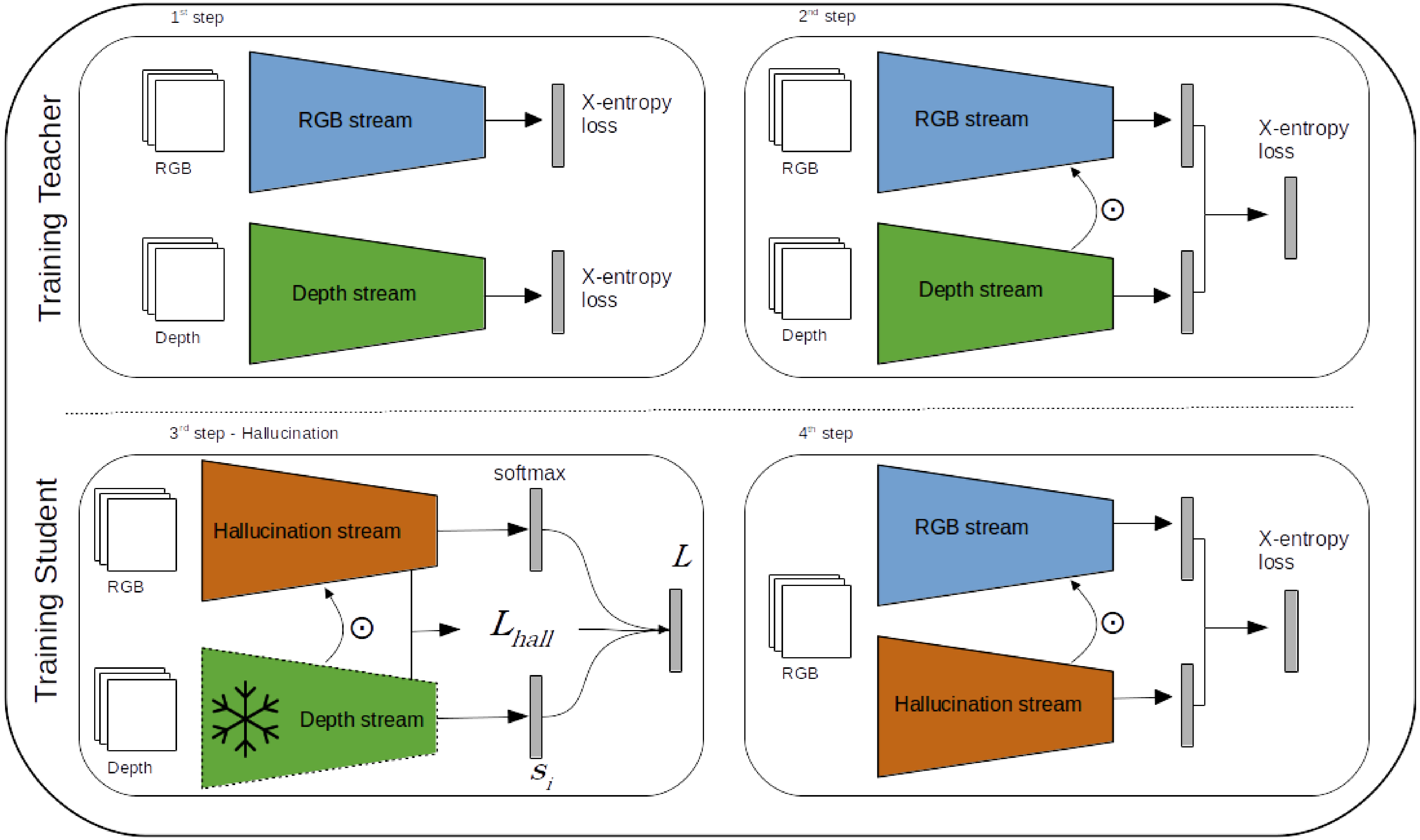
Why hallucinating ?

Depth Perception from Monocular RGB Image

Modality Distillation with Multiple Stream Networks for Action Recognition

N.C. Garcia, P. Morerio, and V. Murino, ECCV 2018

Hallucinating Depth Features from RGB



[1] Feichtenhofer, C., Pinz, A., Wildes, R.P.: Spatiotemporal multiplier networks for video action recognition. In CVPR (2017)

[2] Hoffman, J., Gupta, S., Darrell, T.: Learning with side information through modality hallucination. In CVPR (2016)

[3] Lopez-Paz, D., *et al.* .: Unifying distillation and privileged information. In ICLR (2016)

Knowledge Distillation

- A fact: an ensemble of networks usually performs better than a single network.
- The problem: an ensemble (or a very deep model) may be too heavy for inference in production.
- The idea: Train a single lightweight network to mimic the ensemble of networks (*teacher-student* approach)

Buciluă, C., Caruana, R., Niculescu-Mizil, A.. Model compression. *12th ACM SIGKDD 2006*
Ba, J., Caruana, R.. Do deep nets really need to be deep? *NIPS 2014*

Knowledge Distillation

Data & label: (x, y). Temperature T > 0.

1. Learn Teacher → Ensemble of networks on input pairs (x_i, y_i)
2. Compute Teacher's soft labels
   a. $s_i = \varsigma\,(\text{Teacher}(x_i)\,/\,T)$, ς softmax operator
3. Learn Student → lightweight network using (x_i, y_i) and (x_i, s_i)

What if the ensemble learns from multiple modalities, but the Student network can learn from one only?

Hinton, G., Vinyals, O., Dean, J. "Distilling the knowledge in a neural network.", NIPS 2014 Deep Learning Workshop

Privileged Information

Hypothesis:

- Having access to a Teacher that considers additional information, x^*, together with the pair (data, label) = (x, y),
- and assuming that x^* is not available at test time.

The question is:

- How to leverage the additional information x^* at training time to build a better model that will have access only to x in testing.

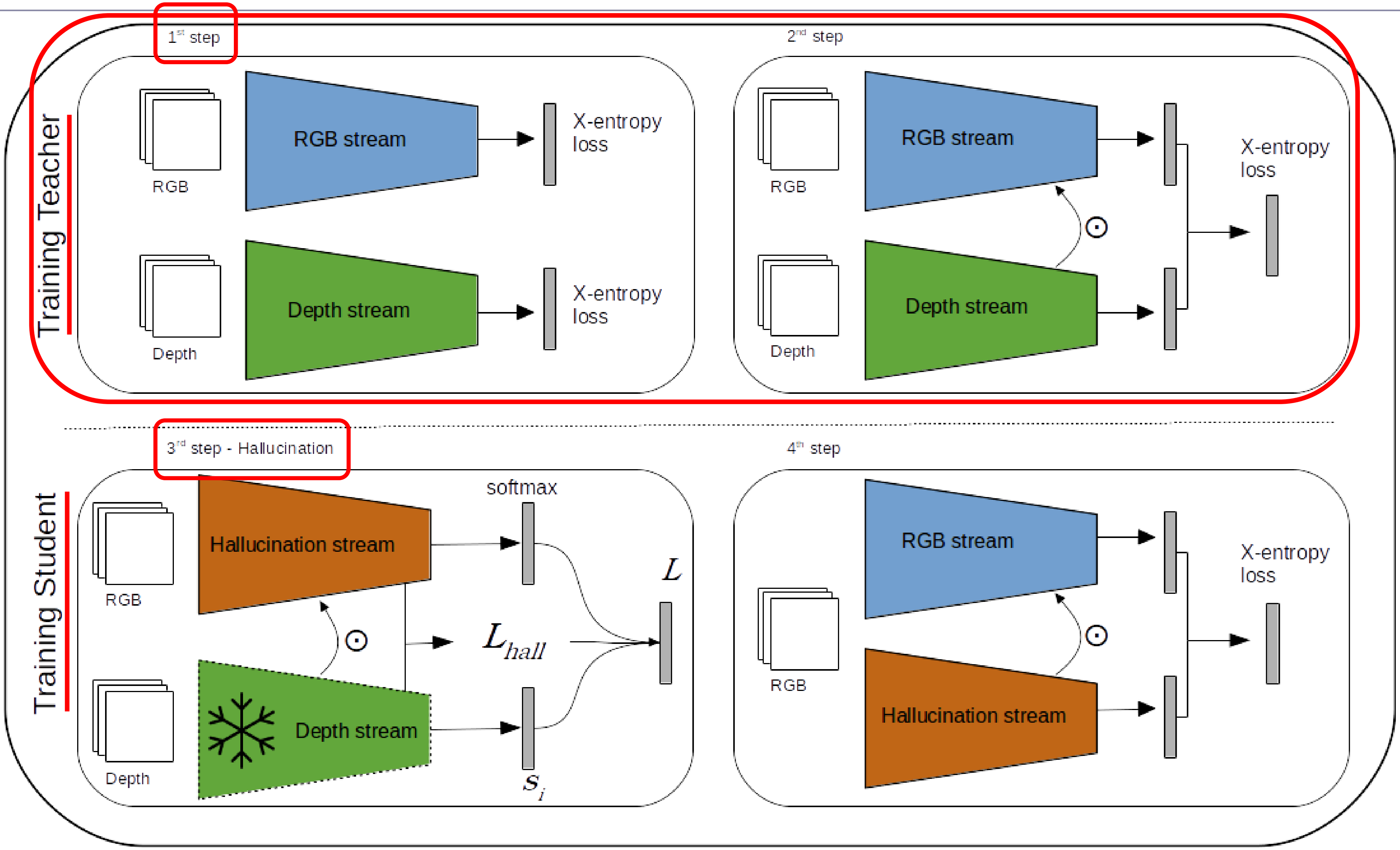
Vapnik, V., Vashist, A. (2009). A new learning paradigm: Learning using privileged information. *Neural networks*, *22*(5-6), 544-557.

Privileged Information and Distillation
Generalized Distillation

- "Machines-teaching-machines" paradigm.
- 3 steps that are common to KD and PI.
- Consider (x, x^*, y):
  1. Learn *Teacher* network on (x^*, y)
  2. Compute Teacher's soft labels as s = ς (Teacher(x^*) / T), ς softmax operator
  3. Learn *Student* network using (x, y) and (x^*, s)
- If $x^* = x$ and the Teacher is bigger than Student network, we are in a Distillation framework.
- If x^* is additional information, we are in a Privileged Information framework.

Lopez-Paz, D., Bottou, L., Schölkopf, B., & Vapnik, V. (2015). Unifying distillation and privileged information. ICLR 2016.

Hallucinating Depth Features from RGB



[1] Feichtenhofer, C., Pinz, A., Wildes, R.P.: Spatiotemporal multiplier networks for video action recognition. In CVPR (2017)

[2] Hoffman, J., Gupta, S., Darrell, T.: Learning with side information through modality hallucination. In CVPR (2016)

[3] Lopez-Paz, D., *et al.* .: Unifying distillation and privileged information. In ICLR (2016)

Losses

- The 1st step refers to the separate (pre-)training of depth and RGB streams with standard cross entropy classification loss.
- The 2nd step represents the actual learning of the Teacher (depth) network
  - Both streams are initialized with the respective weights from step 1 and trained jointly with a cross-entropy loss as a traditional two-stream model, using RGB and depth data.

- We used a similar connection mechanism ($\odot$) between the 2 networks as in *Feichtenhofer et al.*: it is actually implemented at the four convolutional layers of the Resnet-50 model, aiming at learning better spatiotemporal representations

Feichtenhofer, C., Pinz, A., Wildes, R.P. *Spatiotemporal multiplier networks for video action recognition.* CVPR 2017.

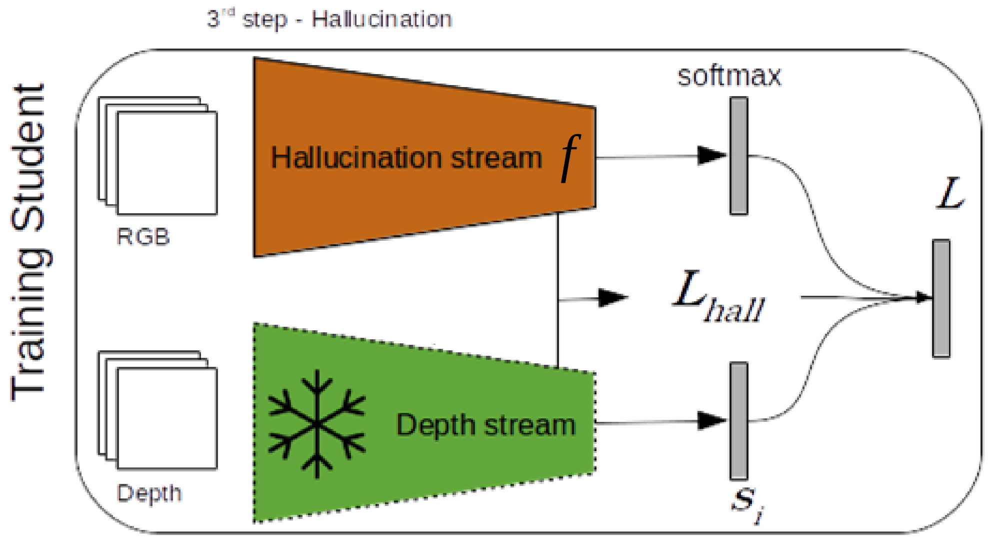
Losses

- 3rd step: learning the student (hallucination) network using Feature (*hallucination*) Loss + Distillation loss
  - *Feature loss:* To align the features of Hallucination Stream with the real Depth Stream.

$$L_{hal}(l) = \lambda_l \|\sigma(A_l^d) - \sigma(A_l^h)\|_2^2$$

where σ is the sigmoid function, and A_l's are the l-th layer activations of depth (*d*) and hallucination (*h*) networks.

- This Euclidean loss forces both activation maps to be similar.



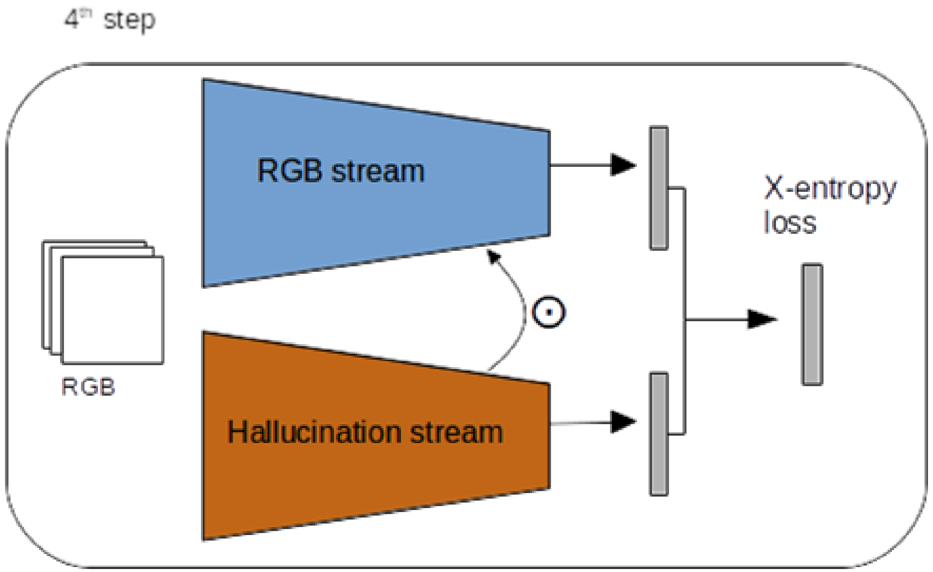
- *Distillation loss:* To align the <u>predictions</u> of Hallucination Stream with the real Depth Stream. The Generalized Distillation Loss is:

$$L_{GD}(i) = (1-\lambda)\ell(y_i, \varsigma(f(x_i))) + \lambda\ell(s_i, \varsigma(f(x_i))),\ \lambda \in [0,1]\ f_s \in \mathcal{F}_s,$$

where $\quad s_i = \varsigma(f_t(x_i)/T),\ T > 0.$

1st term: uses the ground truth labels, y_i
2nd term: uses the soft targets provided by teacher, s_i, and ς is the softmax function

- The final loss is: $L = (1-\alpha)L_{GD} + \alpha L_{hall}$, $\alpha \in [0,1]$



- The 4th and last step refers to a fine-tuning step and also represents the test setup of our model:
  - the hallucination stream is initialized from the respective weights from 3rd step, and the RGB stream with the respective weights from the 2nd step

Training details

- λ, α: balancing the ground truth and soft labels highly depends on the performance of the teacher network.
  - In our experiments we used $\lambda = 0.5$ and $\alpha = 0.5$.
- We used Resnet-50 for all networks.
  - Augmented with 1D temporal convolutions, that span over 5 frames.
- Initialized with ImageNet weights.
- Trained with SGD until validation accuracy reaches a plateau.

Datasets

- NTU RGB+D
  - This is the largest public dataset for multimodal video action recognition.
  - Composed by 56,880 videos, available in four modalities: RGB, depth sequences, infrared frames, and 3D skeleton data of 25 joints.
  - Acquired with a Kinect v2 sensor in 80 different viewpoints, and includes 40 subjects performing 60 distinct actions.
  - We follow the two evaluation protocols originally proposed in Shahroudy et al., which are cross-subject and cross-view.
  - As in the original paper, we use about 5% of the training data as validation set for both protocols, in order to fix the values of parameters and T.
- In this work, we use only RGB and depth data.

Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+ d: A large scale dataset for 3d human activity analysis. CVPR 2016

Datasets

- UWA3DII
  - This dataset consists of 1075 samples of RGB, depth and skeleton sequences.
  - It features 10 subjects performing 30 actions captured in 5 different views.

- Northwestern-UCLA
  - Similarly to the other datasets, it provides RGB, depth and skeleton sequences for 1475 samples.
  - It features 10 subjects performing 10 actions captured in 3 different views.

Comparison with State of the Art

Method	Test Mods.	NTU (p1)	NTU (p2)	UWA3DII	NW-UCLA	
Luo [17]	Depth	66.2%	-	-	-	×
Luo [17]	RGB	56.0%	-	-	-	×
Rahmani [22]	RGB	-	-	67.4%	78.1%	×
HOG-2 [19]	Depth	32.4%	22.3%	-	-	△
Action Tube [7]	RGB	-	-	37.0%	61.5%	△
Ours - depth, step 1	**Depth**	**70.44%**	**75.16%**	**75.28%**	**72.38%**	△
Ours - RGB, step 1	**RGB**	**66.52%**	**71.39%**	**63.67%**	**85.22%**	△
Deep RNN [23]	Joints	56.3%	64.1%	-	-	△
Deep LSTM [23]	Joints	60.7%	67.3%	-	-	△
Sharoudy [23]	Joints	62.93%	70.27%	-	-	△
Kim [26]	Joints	74.3%	83.1%	-	-	△
Sharoudy [24]	RGB+D	74.86%	-	-	-	△
Liu [14]	RGB+D	77.5%	84.5%	-	-	△
Rahmani [20]	Depth+Joints	75.2	83.1	84.2%	-	△
Ours - step 2	**RGB+D**	**79.73%**	**81.43%**	**79.66%**	**88.87%**	△
Hoffman *et al.* [11]	RGB	64.64%	-	66.67%	83.30%	□
Ours - step 3	**RGB**	**71.93%**	**74.10%**	**71.54%**	**76.30%**	□
Ours - step 4	**RGB**	**73.42%**	**77.21%**	**73.23%**	**86.72%**	□

cross-subject *cross-view*

Table 3. Classification accuracies and comparisons with the state of the art. Performances referred to the several steps of our approach (ours) are highlighted in bold. × refers to comparisons with unsupervised learning methods. △ refers to supervised methods: here train and test modalities coincide. □ refers to privileged information methods: here training exploits RGB+D data, while test relies on RGB data only. The 3rd column refers to cross-subject and the 4th to the cross-view evaluation protocols on the NTU dataset. The results reported on the other two datasets are for the cross-view protocol.

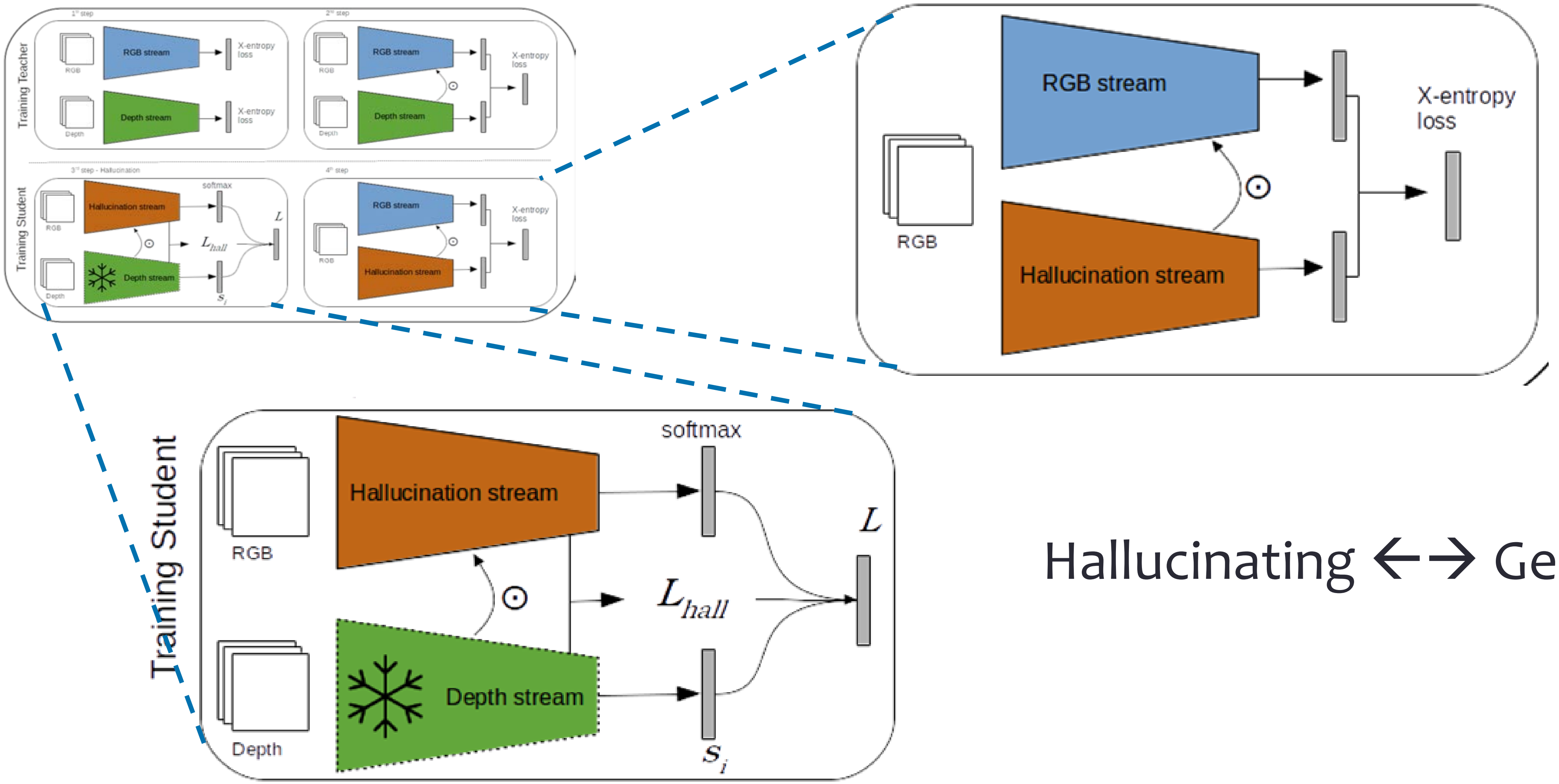
Remarks

- For almost all datasets, Depth outperforms RGB in both Cross-View and Cross-Subject: it makes sense to consider Depth as Teacher modality.
    - Although weak teachers might also improve students, see “Revisiting Knowledge Distillation via Label Smoothing Regularization” by Yuan *et al.* @ CVPR2020
- Network trained with distillation, with RGB as input (Hallucination, step 3) outperforms original (step 1) RGB network.
    - Means that Distillation is indeed providing additional knowledge / regularization effect.
- Still, RGB network + Hallucination achieves the best result (Step 4).
- Hallucination network allows to deal with noisy (depth) data
- This approach can be used in any order of modality (e.g., RGB as Teacher, but it’s less performing)

Learning with privileged information via adversarial discriminative modality distillation

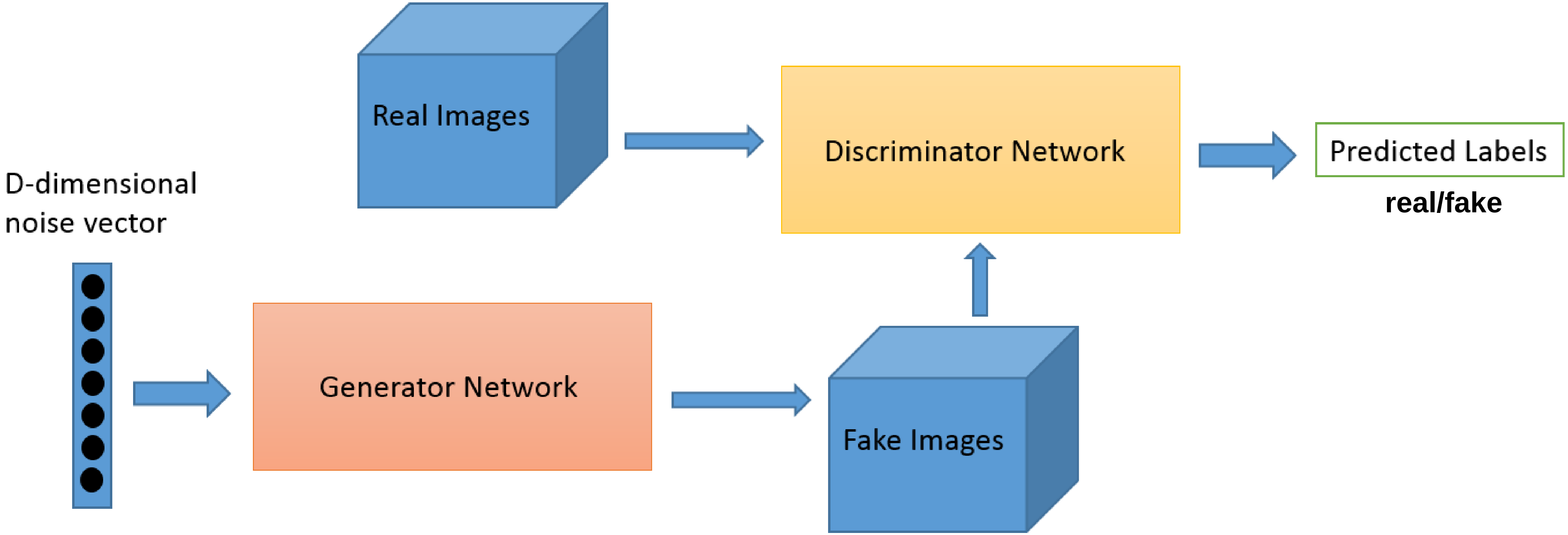
N.C. Garcia, P. Morerio, and V. Murino, IEEE TPAMI (2019)

Hallucinating Depth Features from RGB



Hallucinating ←→ Generating

GAN-based approach



https://skymind.ai/wiki/generative-adversarial-network-gan
Goodfellow et al. "Generative adversarial nets." Advances in Neural Information Processing Systems NIPS 2014.

Learning with privileged information

An adversarial approach

- Adversarial learning strategy to learn the hallucination network.
- The hallucination network plays an adversarial game with a discriminator.

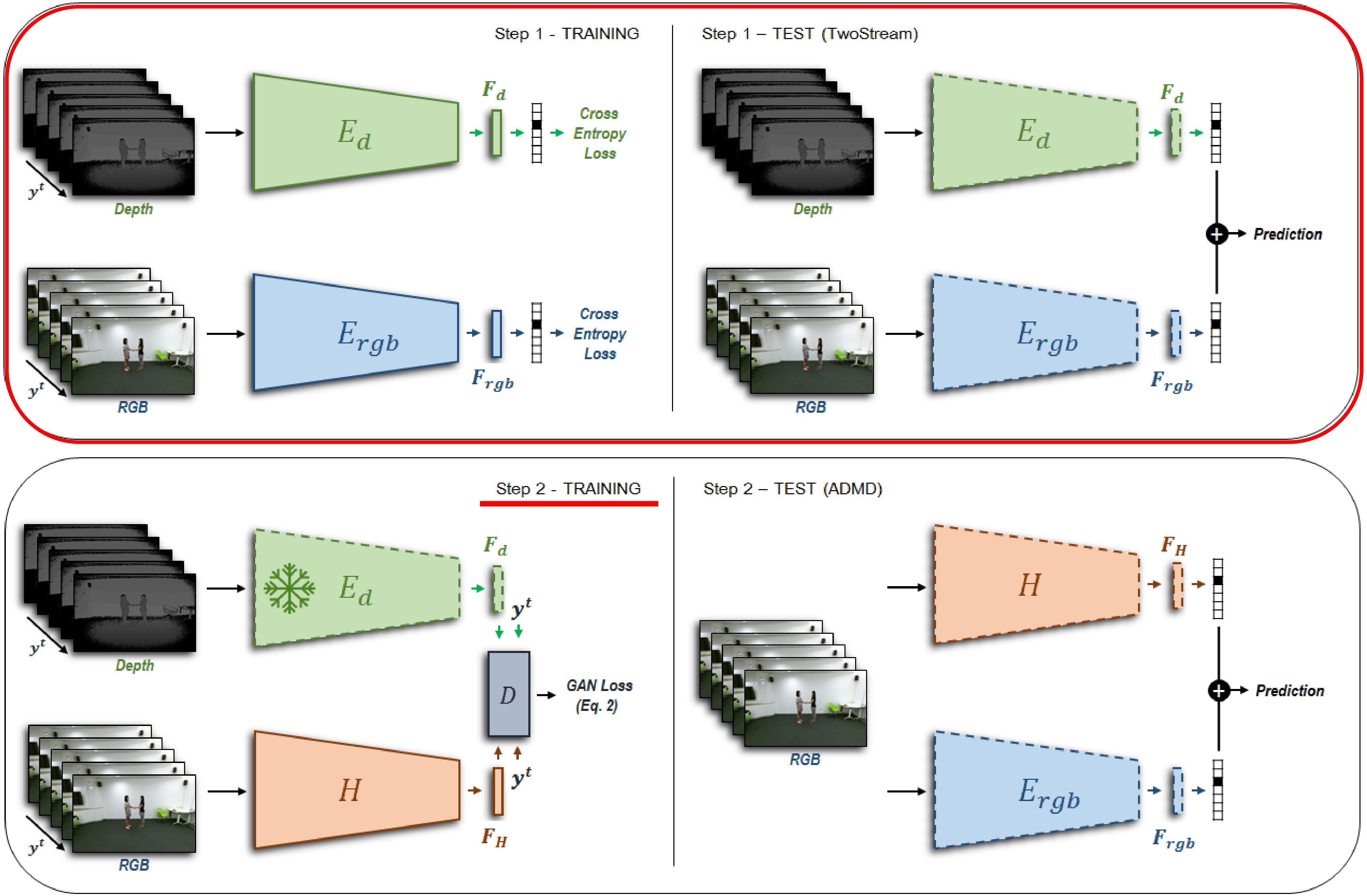
- The discriminator's job is to distinguish between true and hallucinated *features*.
- The hallucination network's job is to fool the discriminator.

- Tasks: video action recognition and object recognition.

Why Adversarial Learning ?

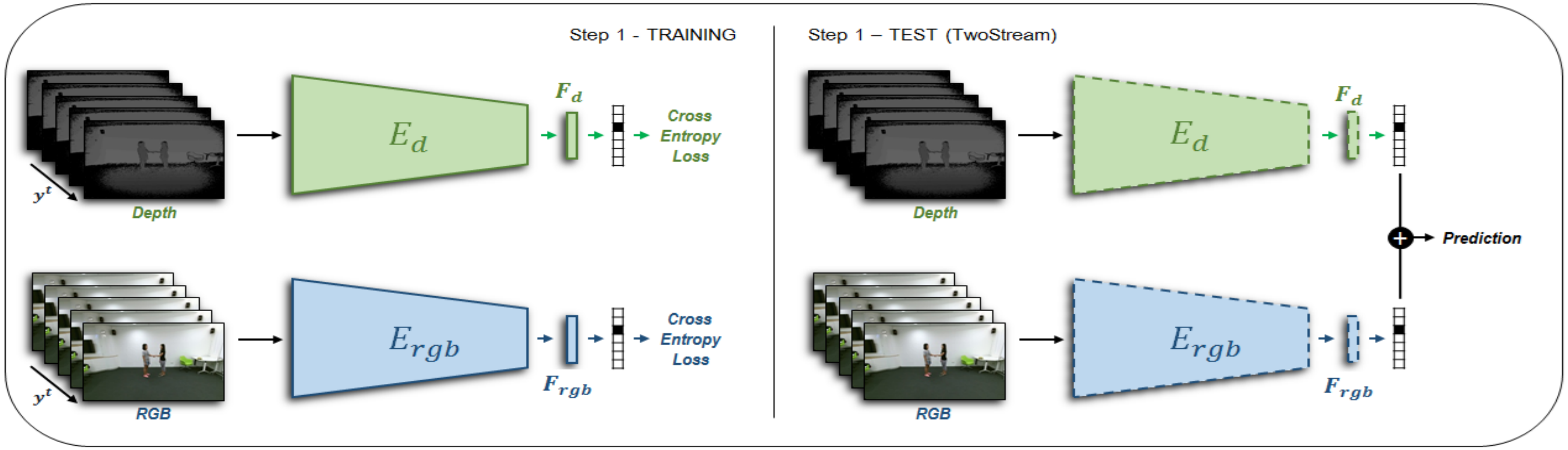
- Using adversarial learning allows for more flexibility.
  - For example, balancing the different losses (Euclidean distance of features and Distillation) can be difficult and certainly varies for different tasks.
- Two-in-one: *align representations and train the classifier* in one objective.
- It provides a mechanism to detect if the modality is too noisy, so we can switch to using the hallucination network.
- It is agnostic regarding the pair of modalities used, being suitable beyond RGB and depth data.
- Thanks to the discriminator design, which includes an auxiliary classification task, our method is able to transfer the discriminative capability from a teacher (depth) network to a student (hallucination) network, up to a full recovery of the teacher’s accuracy.

The adversarial strategy

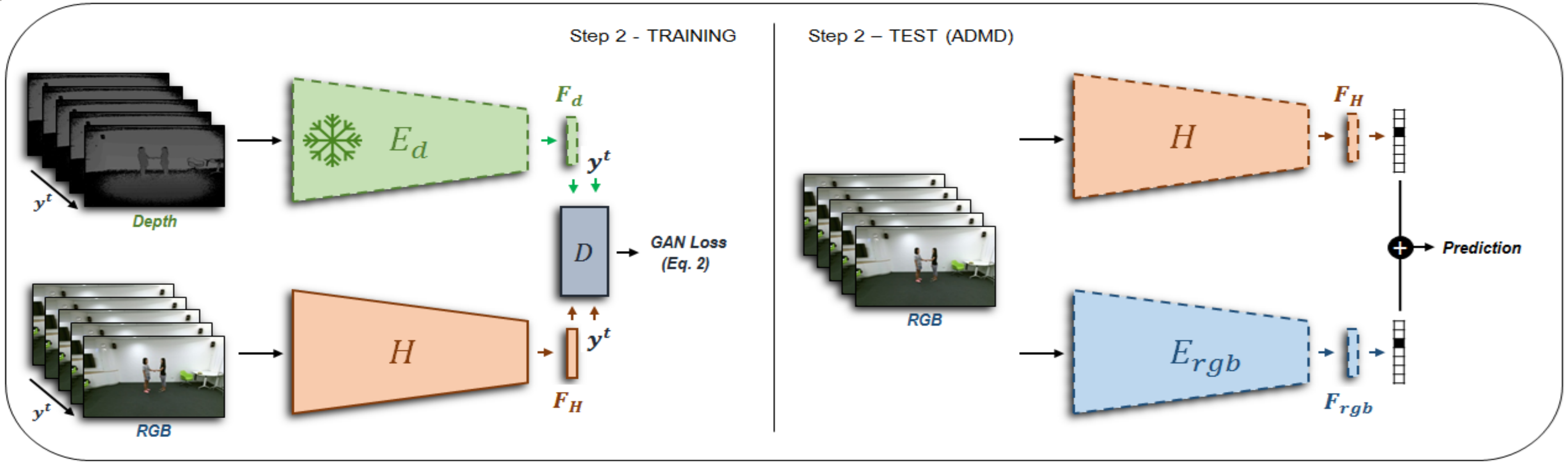


Training

- Step 1: separate training of RGB and depth networks with standard cross-entropy loss.
- At test time the raw predictions (logits) of the two separate streams are simply averaged, boosting the recognition performance.



Training



- <u>Step 2</u> refers to the adversarial training.
  - The Generator role is played by the Hallucination network (*H*), and the "real" target is provided by the Depth network, which is frozen.
  - Input to the Discriminator (*D*): concatenation of the feature vector, and the relative (temporal) position (y^t) of the frame in the video.
  - The discriminator also features the additional classification task of assigning samples to the correct class.
- At test time, predictions from the RGB and the hallucination streams are fused.

Why concatenate the relative position of frame y^t

- Each networks' input is a set of frames.
- Each network is composed by 2D and 1D temporal convolutions.
- The output is a prediction vector for each input frame.
- The first frame and corresponding feature vector might be very different from the last frame / feature vector, even though the prediction should be the same.
- We concatenate the relative position to ground the generator to a position in time, e.g., $y^t = [00100]$ indicates that this frame is sampled from the middle of the clip.

Discriminator task

- **To discriminate** between real depth features and hallucinated features, **and to classify the set of frames.**
- The target $\hat{y}$ and the objective function are defined as:

$$\hat{y} = \begin{cases} [zeros(C)\,||\,1], & \text{for } x_{rgb} \\ [y_i\,||\,0], & \text{for } x_d \end{cases}$$

being y_i the C-dimensional one-hot encoding of the true class label and C is the number of classes

$$\boxed{\begin{aligned} \min_{\theta_D}\max_{\theta_H} \ell = & \ \mathbb{E}_{(x_i,y_i)\sim(X_{rgb},Y)}\ \mathcal{L}(\ D(H(x_i)||y^t),\ \hat{y}_i) \\ & + \mathbb{E}_{(x_i,y_i)\sim(X_d,Y)}\ \mathcal{L}(\ D(E_d(x_i)||y^t),\ \hat{y}_i) \end{aligned}}$$

being $\mathcal{L}$ the cross-entropy

Training details

- Both networks are Resnet-50, augmented with 1D temporal convolutions.
- The input to the Discriminator is obtained from the last feature map of a Resnet-50 [7x7x2048], after pooling and a convolutional layer, to obtain a final vector of size 128.
- The architecture of the Discriminator varies: 3 fully connected (FC) layers for action recognition tasks, and 5 FC for object recognition.
- All networks are trained with Adam, *lr*=0.001, from ImageNet checkpoints.

Experiments
Object Recognition

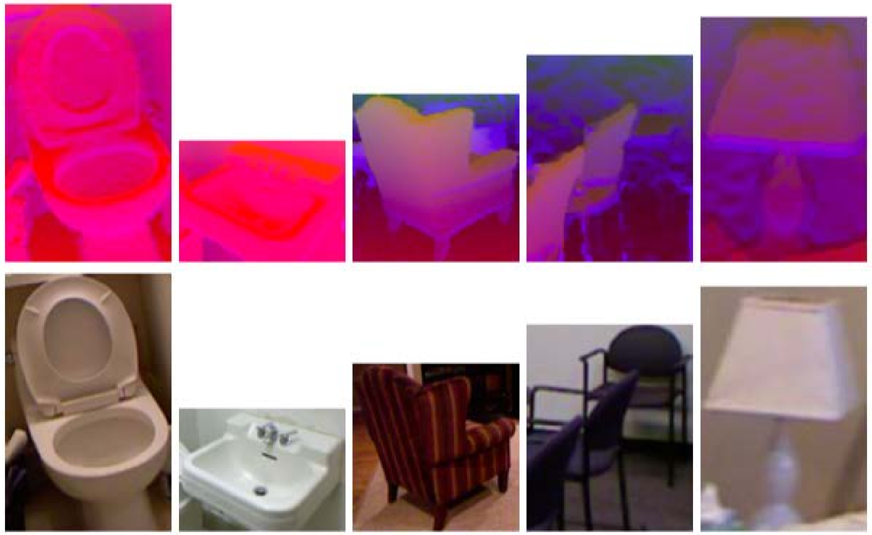


Fig. 5. Examples of RGB and depth frames from the NYUD (RGB-D) dataset.

Object Recognition

Method	Trained on	Tested on	Accuracy
Depth alone	Depth	Depth	40.19%
RGB alone	RGB	RGB	52.90%
RGB ensemble	RGB	RGB	54.14%
Two-stream (average preds.)	RGB+D	RGB+D	57.39%
ModDrop [22]	RGB+D	RGB+D	58.93%
ModDrop [22]	RGB+D	RGB	53.73%
Autoencoder	RGB+D	RGB	50.52%
FCRN [23] depth estimation	RGB+D	RGB	50.23%
Garcia *et al.*	RGB+D	RGB	55.94%
Ours (ADMD)	RGB+D	RGB	**57.52%**

Experiments
Action Recognition

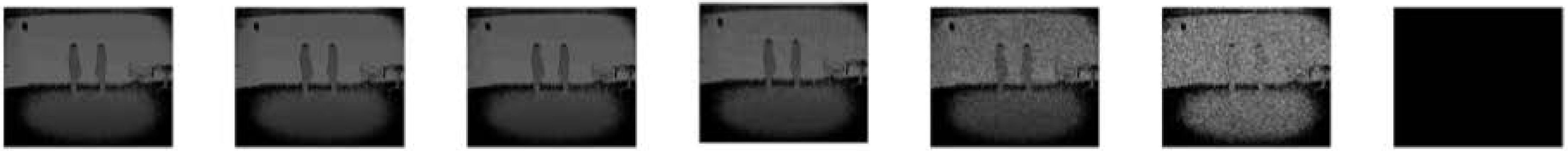
TABLE 4
Classification accuracies and comparisons with the state of the art for video action recognition. Performances referred to the several steps of our approach (ours) are highlighted in bold. × refers to comparisons with unsupervised learning methods. △ refers to supervised methods: here train and test modalities coincide. □ refers to privileged information methods: here training exploits RGB+D data, while test relies on RGB data only. The 4th column refers to cross-subject and the 5th to the cross-view evaluation protocols on the NTU dataset. The results reported on the other two datasets are for the cross-view protocol.

#	Method	Test Mods.	NTU (p1)	NTU (p2)	NW-UCLA	
1	Luo [58]	Depth	66.2%	-	-	
2	Luo [58]	RGB	56.0%	-	-	×
3	Rahmani [59]	RGB	-	-	78.1%	
4	HOG-2 [60]	Depth	32.4%	22.3%	-	
5	Action Tube [61]	RGB	-	-	61.5%	
6	Depth stream [11]	Depth	70.44%	75.16%	72.38%	
7	**ADMD** - Depth stream	Depth	70.53%	76.47%	-	
8	**ADMD** - Depth stream w/ bott.	Depth	71.87%	75.32%	71.09%	
9	[11] - RGB stream	RGB	66.52%	80.01%	85.22%	
10	**ADMD** - RGB stream	RGB	67.95%	80.01%	85.87%	
11	Deep RNN [16]	Joints	56.3%	64.1%	-	△
12	Deep LSTM [16]	Joints	60.7%	67.3%	-	
13	Sharoudy [16]	Joints	62.93%	70.27%	-	
14	Kim [62]	Joints	74.3%	83.1%	-	
15	Sharoudy [5]	RGB+D	74.86%	-	-	
16	Liu [6]	RGB+D	77.5%	84.5%	-	
17	Rahmani [63]	Depth+Joints	75.2	83.1	-	
18	Two-stream, step 2 [11]	RGB+D	79.73%	81.43%	88.87%	
19	**ADMD** - Two-stream (no finetune)	**RGB+D**	**77.74%**	**85.49%**	**89.93%**	
20	Hoffman *et al.* [10]	RGB	64.64%	-	83.30%	
21	Luo *et al.* [21]	RGB	89.50%	-	-	
22	Hallucination model, step 3 [11]	RGB	71.93%	74.10%	76.30%	
23	Hallucination model, step 4 [11]	RGB	73.42%	77.21%	86.72%	□
24	**ADMD** - Hall. stream alone	**RGB**	**67.57%**	**71.80%**	**83.94%**	
25	**ADMD** - Hall. two-stream model	**RGB**	**73.11%**	**81.50%**	**91.64%**	

Robustness to depth noise

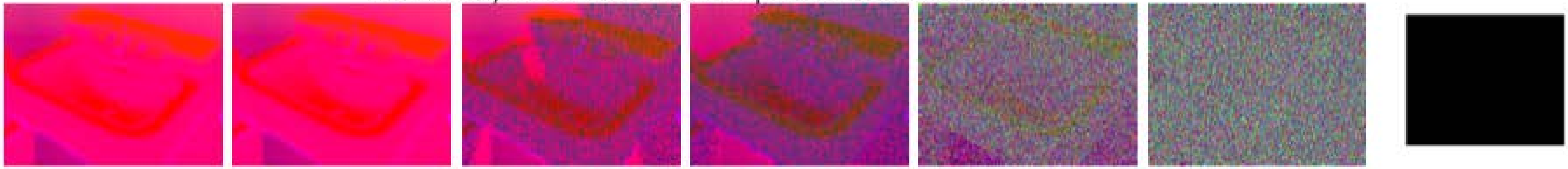
Accuracy values for the two-stream model trained on RGB and depth, and tested with RGB and noisy depth data.

NTU RGB+D action dataset - ADMD performance is 81.50%.



σ^2	*no noise*	10^{-3}	10^{-2}	10^{-1}	10^{0}	10^{1}	*void*
Two-stream	85.49%	85.52%	82.05%	68.99%	2.16%	3.35%	8.55%

NYUD object dataset - ADMD performance is 57.52%.



σ^2	*no noise*	10^{-3}	10^{-2}	10^{-1}	10^{0}	10^{1}	*void*
Two-stream	58.73%	58.68%	58.23%	57.18%	48.27%	28.40%	47.44%
ModDrop [20]	58.93%	58.89%	58.56%	57.49%	48.90%	25.95%	47.86%

Detecting noisy inputs

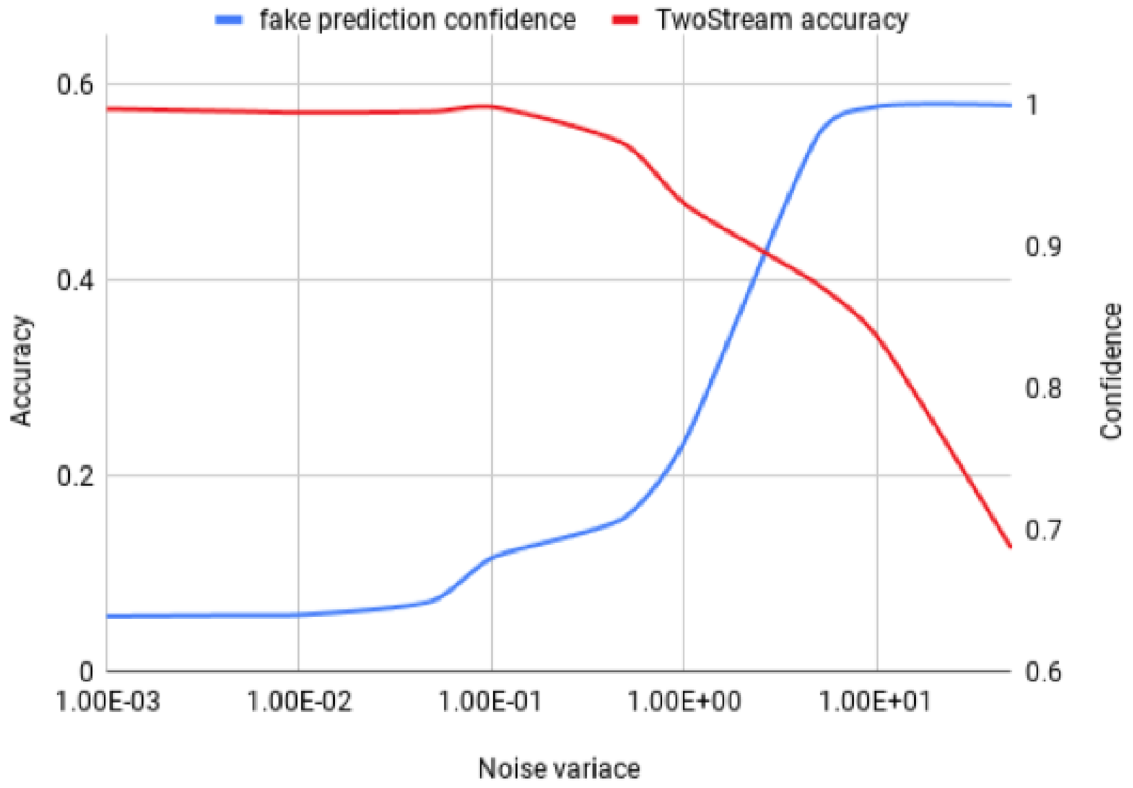


Fig. 6. Discriminator confidence at predicting 'fake' label as a function of noise in the depth frames. The more corrupted the frame, the more confident D, and the lower the accuracy of the Two-stream model (NYUD dataset).

Insights from Multimodal Learning and Distillation

- If a modality is missing and the task is not pixel-level, one can still bring some of the performance offered by that modality without actually predicting the modality itself, but just the features.
  - Autoencoder 50.52% vs. Ours 57.52%
- Adversarial Learning (Real / Synthetic) is not enough to learn good features: the auxiliary classification task is important.
- The fact that [RGB + Hallucination stream] ensemble outperforms [RGB + RGB] indicates that Distillation is not only a way to regularize the learning, but also transfers knowledge.

Conclusions

- Multimodal Learning poses interesting and unsolved challenges
  - Learning representations from multiple modalities while leveraging the potential of each one efficiently.
  - Applying the models in real scenarios, *e.g.*, missing or noisy modalities.
- KD and PI provide a framework for learning using multiple streams of information.
- The test-time modality network can be enriched with information coming from other modalities / networks
  - either by having a hallucination network providing the missing predictions
  - or by simply using the extra pairs (x*, s) to enlarge the dataset.

Conclusions

- KD is a regularization method, but a special one:
  - it also serves to provide additional, useful information to the student network.

- It is related to the idea of noisy labels or label augmentation

- Embodies the idea that using hard labels is not always optimal.

Acknowledgments

Nuno C. Garcia

Pietro Morerio

… and thanks for the attention

- Code:
  https://github.com/ncgarcia/modality-distillation
  https://github.com/ncgarcia/admd