Modality Distillation with Multiple Stream Networks for Action Recognition

Nuno C. Garcia[1,2]([envelope]) [ID], Pietro Morerio[1] [ID], and Vittorio Murino[1,3] [ID]

[1] Istituto Italiano di Tecnologia, Genoa, Italy
{nuno.garcia,pietro.morerio,vittorio.murino}@iit.it
[2] Università degli Studi di Genova, Genoa, Italy
[3] Università di Verona, Verona, Italy

Abstract. Diverse input data modalities can provide complementary cues for several tasks, usually leading to more robust algorithms and better performance. However, while a (training) dataset could be accurately designed to include a variety of sensory inputs, it is often the case that not all modalities are available in real life (testing) scenarios, where a model has to be deployed. This raises the challenge of how to learn robust representations leveraging multimodal data in the training stage, while considering limitations at test time, such as noisy or missing modalities. This paper presents a new approach for multimodal video action recognition, developed within the unified frameworks of distillation and privileged information, named generalized distillation. Particularly, we consider the case of learning representations from depth and RGB videos, while relying on RGB data only at test time. We propose a new approach to train an hallucination network that learns to distill depth features through multiplicative connections of spatiotemporal representations, leveraging soft labels and hard labels, as well as distance between feature maps. We report state-of-the-art results on video action classification on the largest multimodal dataset available for this task, the NTU RGB+D, as well as on the UWA3DII and Northwestern-UCLA.

Keywords: Action recognition · Deep multimodal learning Distillation · Privileged information

1 Introduction

Imagine you have a large multimodal dataset to train a deep learning model on, comprising for example RGB video sequences, depth maps, infrared, and skeleton joints data. However, at test time, the trained model may be deployed in scenarios where not all of these modalities are available. For example, most of the cameras capture RGB only, which is the most common and cheapest available data modality.

Considering this limitation, we would like to answer the following: what is the best way of using all data available in order to learn robust representations to be exploited when there are missing modalities at test time? In other words, is there any added value in training a model by exploiting more data modalities, even if only one can be used at test time? Unsurprisingly, the simplest and most commonly adopted solution consists in training the model using only the modality in which it will be tested. However, a more interesting alternative is trying to exploit the potential of the available data and train the model using all modalities, realizing, however, that not all of them will be accessible at test time. This learning paradigm, i.e., when the model is trained using extra information, is generally known as *learning with privileged information* [30] or *learning with side information* [11].

In this work, we propose a multimodal stream framework that learns from different data modalities and can be deployed and tested on a subset of these. We design a model able to learn from RGB *and* depth video sequences, but due to its general structure, it can also be used to manage whatever combination of other modalities as well. To show its potential, we evaluate the performance on the task of video action recognition. In this context, we introduce a new learning paradigm, depicted in Fig. 1, to *distill* the information conveyed by depth into an *hallucination* network, which is meant to "mimic" the missing stream at test time. Distillation [1,10] refers to any training procedure where knowledge is transferred from a previously trained complex model to a simpler one. Our learning procedure introduces a new loss function which is inspired to the *generalized distillation* framework [15], that formally unifies distillation and privileged information learning theories.

Our model is inspired to the two-stream network introduced by Simonyan and Zisserman [25], which has been notably successful in the traditional setting for video action recognition task [2,5]. Differently from previous works, we use multimodal data, deploying one stream for each modality (RGB and depth in our case), and use it in the framework of privileged information. Another inspiring work for our framework is [11], which proposes an hallucination network to learn with side information. We build on this idea, extending it by devising a new mechanism to *learn* and *use* such hallucination stream through a more general loss function and inter-stream connections.

To summarize, the main contributions of this paper are the following:

– we propose a new multimodal stream network architecture able to exploit multiple data modalities at training while using only one at test time;
– we introduce a new paradigm to learn an hallucination network within a novel two-stream model;
– in this context, we have implemented an inter-stream connection mechanism to improve the learning process of the hallucination network, and designed a more general loss function, based on the generalized distillation framework;
– we report state-of-the-art results – in the privileged information scenario – on the largest multimodal dataset for video action recognition, the NTU RGB+D [23], and on two other smaller ones, the UWA3DII [21] and the Northwestern-UCLA [33].

The rest of the paper is organized as follows. Section 2 reviews similar approaches and discusses how they relate to the present work. Section 3 details the proposed architecture and the novel learning paradigm. Section 4 reports the results obtained on the various datasets, including a detailed ablation study performed on the NTU RGB+D dataset and a comparative performance with respect to the state of the art. Finally, we draw conclusions and future research directions in Sect. 5.

2   Related Work

Our work is at the intersection of three topics: privileged information [30], network distillation [1,10], and multimodal video action recognition. However, Lopez-Paz *et al.* [15] noted that privileged information and network distillation are instances of a the same more inclusive theory, called generalized distillation.

Generalized Distillation. Within the generalized distillation framework, our model is both related to the privileged information theory [30], considering that the extra modality (depth, in this case) is only used at training time, and, mostly, to the distillation framework. In fact, the core mechanism that our model uses to learn the hallucination network is derived from a distillation loss. More specifically, the supervision information provided by the teacher network (in this case, the network processing the depth data stream) is distilled into the hallucination network leveraging teacher's soft predictions and hard ground-truth labels in the loss function.

In this context, the closest works to our proposal are [11,16]. Luo *et al.* [16] addressed a similar problem to ours, where the model is first trained on several modalities (RGB, depth, joints and infrared), but tested only in one. The authors propose a graph-based distillation method that is able to distill information from all modalities at training time, while also passing through a validation phase on a subset of modalities. This showed to reach state-of-the-art results in action recognition and action detection tasks. Our work substantially differs from [16] since we benefit from an hallucination mechanism, consisting in an auxiliary network trained using the guidance distilled by the *teacher* network (that processes the depth data stream in our case). This mechanism allows the model to learn to emulate the presence of the missing modality at test time.

The work of Hoffman *et al.* [11] introduced a model to hallucinate depth features from RGB input for object detection task. While the idea of using an hallucination stream is similar to the one thereby presented, the mechanism used to learn it is different. In [11], the authors use an Euclidean loss between depth and hallucinated feature maps, that is part of the total loss along with more than ten classification and localization losses, which makes its effectiveness very dependent on hyperparameter tuning to balance the different values, as the model is trained jointly in one step by optimizing the aforementioned composite loss. Differently, we propose a loss inspired to the distillation framework, that not only uses the Euclidean distance between feature maps, and the one-hot

labels, but also leverages soft predictions from the depth network. Moreover, we encourage the hallucination learning by design, by using cross-stream connections (see Sect. 3). This showed to largely improve the performance of our model with respect to the one-step learning process proposed in [11].

Multimodal Video Action Recognition. Video action recognition has a long and rich field of literature, spanning from classification methods using hand-crafted features [3,13,31,32] to modern deep learning approaches [2,12,28,34], using either RGB-only or various multimodal data. Here, we focus on some of the more relevant works in multimodal video action recognition, including state-of-the-art methods considering the NTU RGB+D dataset, as well as architectures related to our proposed model.

The two-stream model introduced by Simonyan and Zisserman [25] is a landmark on video analysis, and since then has inspired a series of variants that achieved state-of-the-art performance on diverse datasets. This architecture is composed by an RGB and an optical flow stream, which are trained separately, and then fused at the prediction layer. The current state of the art in video action recognition [2] is inspired by such model, featuring 3D convolutions to deal with the temporal dimension, instead of the original 2D ones. In [5], a further variation of the two-stream approach is proposed, which models spatiotemporal features by injecting the motion stream's signal into the residual unit of the appearance stream. The idea of combining the two streams have also been explored previously by the same authors in [6].

Instead, in [24], the authors explore the complementary properties of RGB and depth data, taking the NTU RGB+D dataset as testbed. This work designed a deep autoencoder architecture and a structured sparsity learning machine, and showed to achieve state-of-the-art results for action recognition. Liu *et al.* [14] also use RGB and depth complementary information to devise a method for viewpoint invariant action recognition. Here, dense trajectories from RGB data are first extracted, which are then encoded in viewpoint invariant deep features. The RGB and depth features are then used as a dictionary to predict the test label.

All these previous methods exploited the rich information conveyed by the multimodal data to improve recognition. Our work, instead, proposes a fully convolutional model that exploits RGB and depth data at training time only, and uses exclusively RGB data as input at test time, reaching performance comparable to those utilizing the complete set of modalities in both stages.

3 Generalized Distillation with Multiple Stream Networks

This section describes our approach in terms of its architecture, the losses used to learn the different networks, and the training procedure.
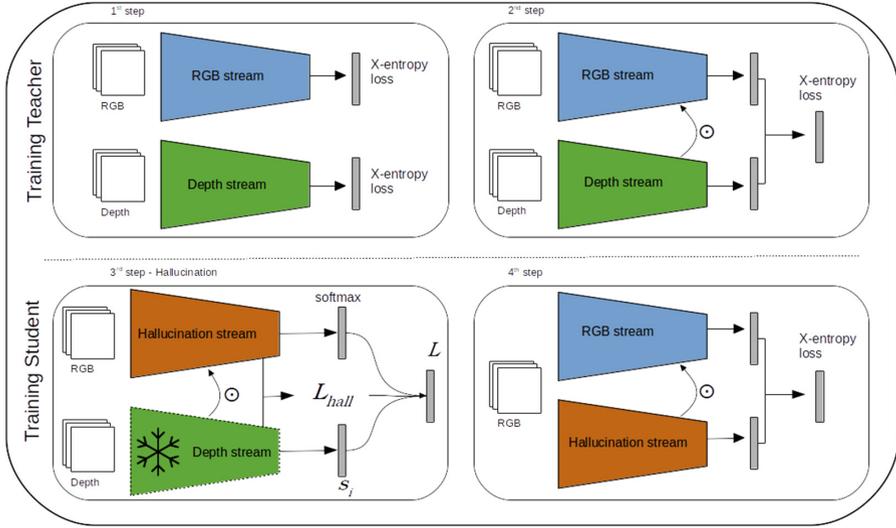
Fig. 1. Training procedure described in Sect. 3.3 (see also text therein). The 1st step refers to the separate (pre-)training of depth and RGB streams with standard cross entropy classification loss, with both streams initialized with ImageNet weights. The 2nd step represents the learning of the teacher 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. The 3rd step represents the learning of the student network: both streams are initialized with the depth stream weights from the previous step, but the actual depth stream is frozen; importantly, the input for the hallucination stream is RGB data; the model is trained using the loss proposed in Eq. 5. 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 previous step, and the RGB stream with the respective weights from the 2nd step; this model is fine-tuned using a cross entropy loss, and importantly, using only RGB data as input for both streams.

3.1   Cross-Stream Multiplier Networks

Typically in two-stream architectures, the two streams are trained separately and the predictions are fused with a late fusion mechanism [5,25]. Such models use as input appearance (RGB) and motion (optical flow) data, which are fed separately into each stream, both in training and testing. Instead, in this paper we use RGB and depth frames as inputs for training, but only RGB at test time, as already discussed (Fig. 1).

We use the ResNet-50-based [8,9] model proposed in [5] as baseline architecture for each stream block of our model. In that paper, Feichtenhofer *et al.* proposed to connect the appearance and motion streams with multiplicative connections at several layers, as opposed to previous models which would only interact at the prediction layer. Such connections are depicted in Fig. 1 with the ⊙ symbol. Figure 2 illustrates this mechanism at a given layer of the

multiple stream architecture, but, in our work, it is actually implemented at the four convolutional layers of the Resnet-50 model. The underlying intuition is that these connections enable the model to learn better spatiotemporal representations, and help to distinguish between identical actions that require the combination of appearance and motion features. Originally, the cross-stream connections consisted in the injection of the motion stream signal into the other stream's residual unit, without affecting the skip path. ResNet's residual units are formally expressed as:

$$\mathbf{x}_{l+1} = f(h(\mathbf{x}_l) + F(\mathbf{x}_l, \mathcal{W}_l)),$$

where $\mathbf{x}_l$ and $\mathbf{x}_{l+1}$ are l-th layer's input and output, respectively, F represents the residual convolutional layers defined by weights $\mathcal{W}_l$, $h(\mathbf{x}_l)$ is an identity mapping and f is a ReLU non-linearity. The cross-streams connections are then defined as

$$\mathbf{x}_{l+1}^a = f(\mathbf{x}_l^a) + F(\mathbf{x}_l^a \odot f(\mathbf{x}_l^m), \mathcal{W}_l),$$

where $\mathbf{x}^a$ and $\mathbf{x}^m$ are the appearance and motion streams, respectively, and $\odot$ is the element-wise multiplication operation. Such mechanism implies a spatial alignment between both feature maps, and therefore between both modalities. This alignment comes for free when using RGB and optical flow, since the latter is computed from the former in a way that spatial arrangement is preserved. However, this is an assumption we can not generally make. For instance, depth and RGB are often captured from different sensors, likely resulting in spatially misaligned frames. We cope with this alignment problem in the method's initialization phase (described in the supplementary material).

Temporal Convolutions. In order to augment the model temporal support, we implement 1D temporal convolutions in the second residual unit of each ResNet layer (as in [5]), as illustrated in Fig. 2. The weights $W_l \in \mathbb{R}^{1 \times 1 \times 3 \times C_l \times C_l}$ are convolutional filters initialized as identity mappings at feature level, and centered in time, and C_l is the number of channels in layer l.

Very recently in [29], the authors explored various network configurations using temporal convolutions, comparing several different combinations for the task of video classification. This work suggests that decoupling 3D convolutions into 2D (spatial) and 1D (temporal) filters is the best setup in action recognition tasks, producing best accuracies. The intuition for the latter setup is that factorizing spatial and temporal convolutions in two consecutive convolutional layers eases training of the spatial and temporal tasks (also in line with [27]).

3.2 Hallucination Stream

We also introduce and learn an hallucination network [11], using a new learning paradigm, loss function and interaction mechanism. The hallucination stream network has the same architecture as the appearance and depth stream models.

This network receives RGB as input, and is trained to "imitate" the depth stream at different levels, *i.e.* at feature and prediction layers. In this paper, we
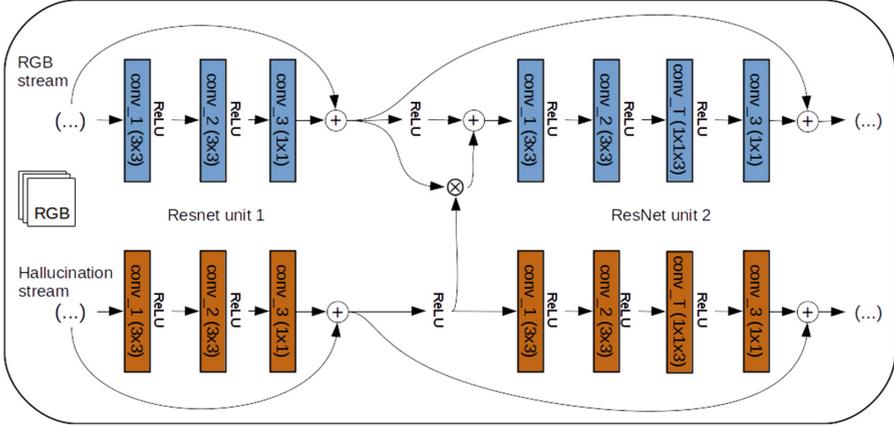
Fig. 2. Detail of the ResNet residual unit, showing the multiplicative connections and temporal convolutions [5]. In our architecture, the signal injection occurs before the 2nd residual unit of each of the four ResNet blocks.

explore several ways to implement such learning paradigm, including both the training procedure and the loss, and how they affect the overall performance of the model.

In [11] it is proposed a regression loss between the hallucination and depth feature maps, defined as:

$$L_{hall}(l) = \lambda_l \|\sigma(A_l^d) - \sigma(A_l^h)\|_2^2, \tag{1}$$

where σ is the sigmoid function, and A_l^d and A_l^h are the l-th layer activations of depth and hallucination network. This Euclidean loss forces both activation maps to be similar. In [11], this loss is weighted along with another ten classification and localization loss terms, making it hard to balance the total loss. One of the main motivations behind the proposed new staged learning paradigm, described in Sect. 3.3, is to avoid the inefficient, heuristic-based tweaking of so many loss weights, aka hyper-parameter tuning.

Instead, we adopt an approach inspired by the generalized distillation framework [15], in which a *student* model $f_s \in \mathcal{F}_s$ distills the representation $f_t \in \mathcal{F}_t$ learned by the *teacher* model. This is formalized as

$$f_s = \arg\min_{f \in \mathcal{F}_s} \frac{1}{n} \sum_{i=1}^{n} L_{GD}(i), n = 1, ..., N \tag{2}$$

where N is the number of examples in the dataset. The generalized distillation loss is so defined as:

$$L_{GD}(i) = (1 - \lambda)\ell(y_i, \varsigma(f(x_i))) + \lambda\ell(s_i, \varsigma(f(x_i))), \ \lambda \in [0, 1] \ f_s \in \mathcal{F}_s, \tag{3}$$

ς is the *softmax* operator and s_i is the soft prediction from the teacher network:

$$s_i = \varsigma(f_t(x_i)/T), \ T > 0. \tag{4}$$

The parameter λ in Eq. 3 allows to tune the loss by giving more importance either to imitating ground-truth hard or soft teacher targets, y_i and s_i, respectively. This mechanism indeed allows the transfer of information from the depth (teacher) to the hallucination (student) network. The temperature parameter T in Eq. 4 allows to smooth the probability vector predicted by the teacher network. The intuition is that such smoothing may expose relations between classes that would not be easily revealed in raw predictions, further facilitating the distillation by the student network f_s.

We suggest that both Euclidean and generalized distillation losses are indeed useful in the learning process. In fact, by encouraging the network to decrease the distance between hallucinated and true depth feature maps, it can help to distill depth information encoded in the generalized distillation loss. Thus, we formalize our final loss function as follows:

$$L = (1 - \alpha)L_{GD} + \alpha L_{hall}, \ \alpha \in [0, 1], \tag{5}$$

where α is a parameter balancing the contributions of the two loss terms during training. The parameters λ, α and T are estimated by utilizing a validation set. The details for their setting are provided in the supplementary material.

In summary, the generalized distillation framework proposes to use the student-teacher framework introduced in the distillation theory to extract knowledge from the privileged information source. We explore this idea by proposing a new learning paradigm to train an hallucination network using privileged information, which we will describe in the next section. In addition to the loss functions introduced above, we also allow the teacher network to share information with the student network by design, through the cross-stream multiplicative connections. We test how all these possibilities affect the model's performance in the experimental section through an extensive ablation study.

3.3 Training Paradigm

In general, the proposed training paradigm, illustrated in Fig. 1, is divided in two core parts: the first part (Step 1 and 2 in the figure) focuses on learning the teacher network f_t, leveraging RGB and depth data (the privileged information in this case); the second part (Step 3 and 4 in the figure) focuses on learning the hallucination network, referred to as student network f_s in the distillation framework, using the general hallucination loss defined in Eq. 5.

The *first training step* consists in training both streams separately, which is a common practice in two-stream architectures. Both depth and appearance streams are trained minimizing cross-entropy, after being initialized with a pre-trained ImageNet model for all experiments. Temporal kernels are initialized as $[0, 1, 0]$, *i.e.* only information on the central frame is used at the beginning - this eventually changes as the training continues. As in [4], depth frames are encoded into color images using a jet colormap.

The *second training step* is still focused on further training the teacher model. Since the model trained in this step has the architecture and capacity of the final

one, and *has access to both modalities*, its performance represents an upper bound for the task we are addressing. This is one of the major differences between our approach and the one used in [11]: by decoupling the teacher learning phase with the hallucination learning, we are able to both learn a better teacher *and* a better student, as we will show in the experimental section.

In the *third training step*, we focus on learning the hallucination network from the teacher model, *i.e.*, the depth stream network just trained. Here, the weights of the depth network are frozen, while receiving in input depth data. Instead, the hallucination network, receiving in input RGB data, is trained with the loss defined in (5), while also receiving feedback from the cross-stream connections from the depth network. We found that this helps the learning process.

In the *fourth and last step*, we carry out fine tuning of the whole model, composed by the RGB and the hallucination streams. This step uses RGB only as input, and it also precisely resembles the setup used at test time. The cross-stream connections inject the hallucinated signal into the appearance RGB stream network, resulting in the multiplication of the hallucinated feature maps and the RGB feature maps. The intuition is that the hallucination network has learned to inform the RGB model where the action is taking place, similarly to what the depth model would do with real depth data.

4   Experiments

4.1   Datasets

We evaluate our method on three datasets, while the ablation study is performed only on the NTU RGB+D dataset. Our model is initialized with ImageNet pre-trained weights and trained and evaluated on the NTU RGB+D dataset. We later fine-tune this model on each of the two smaller datasets for the corresponding evaluation experiments.

NTU RGB+D [23]**.** This is the largest public dataset for multimodal video action recognition. It is composed by 56,880 videos, available in four modalities: RGB videos, depth sequences, infrared frames, and 3D skeleton data of 25 joints. It was 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 [23], 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 select the parameters λ, α and T. In this work, we use only RGB and depth data. The masked depth maps are converted to a three channel map via a jet mapping, as in [4].

UWA3DII [21]**.** This dataset consists on 1075 samples of RGB, depth and skeleton sequences. It features 10 subjects performing 30 actions captured in 5 different views.

Northwestern-UCLA [33]**.** 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.

4.2   Ablation Study

In this section we discuss the results of the experiments carried out to understand the contribution of each part of the model and of the training procedure. Table 1 reports performances at the several training steps, different losses and model configurations.

Table 1. Ablation study. A full set of experiments is provided for the NTU cross-subject evaluation protocol. For cross-view protocol, only the most important results are reported.

#	Method	Test modality	Loss	Cross-Subject	Cross-View
1	Ours - step 1, depth stream	Depth	x-entr	70.44%	75.16%
2	Ours - step 1, RGB stream	RGB	x-entr	66.52%	71.39%
3	Hoffman [11] w/o connections	RGB	Equation (1)	64.64%	–
4	Hoffman [11] w/o connections	RGB	Equation (3)	68.60%	–
5	Hoffman [11] w/o connections	RGB	Equation (5)	70.70%	–
6	Ours - step 2, depth stream	Depth	x-entr	71.09%	77.30%
7	Ours - step 2, RGB stream	RGB	x-entr	66.68%	56.26%
8	Ours - step 2	RGB & Depth	x-entr	**79.73**%	81.43%
9	Ours - step 2 w/o connections	RGB & Depth	x-entr	78.27%	**82.11**%
10	Ours - step 3 w/o connections	RGB (*hall*)	Equation (1)	69.93%	70.64%
11	Ours - step 3 w/ connections	RGB (*hall*)	Equation (1)	70.47%	–
12	Ours - step 3 w/ connections	RGB (*hall*)	Equation (3)	71.52%	–
13	Ours - step 3 w/ connections	RGB (*hall*)	Equation (5)	71.93%	74.10%
14	Ours - step 3 w/o connections	RGB (*hall*)	Equation (5)	71.10%	–
15	**Ours - step 4**	**RGB**	x-entr	**73.42**%	**77.21**%

Rows #1 and #2 refer to the first training step, where depth and RGB streams are trained separately. We note that the depth stream network provides better performance with respect to the RGB one, as expected.

The second part of the table (Rows #3–5) shows the results using Hoffman *et al.*'s method [11] - *i.e.*, adopting a model initialized with the pre-trained networks from the first training step, and the hallucination network initialized using the depth network. Row #3 refers to the original paper [11] (i.e., using the loss L_{hall}, Eq. 1), and rows #4 and #5 refer to the training using the proposed losses L_{GD} and L, in Eqs. 3 and 5, respectively. It can be noticed that the accuracies achieved using the proposed loss functions overcome that obtained in [11] by a significant margin (about 6% in the case of the total loss L).

The third part of the table (rows #6–9) reports performances after the training step 2. Rows #6 and #7 refer to the accuracy provided by depth and RGB stream networks belonging to the model of row #8, taken individually. The final model constitutes the upper bound for our hallucination model, since it uses RGB and depth for training and testing. Performances obtained by the model

in row #8 and #9, with and without cross-stream connections, respectively, are the highest in absolute since using both modalities (around 78–79% for cross-subject and 81–82% for cross-view protocols, respectively), largely outperforming the accuracies obtained using only one modality (in rows #6 and #7).

The *fourth* part of the table (rows #10–14) shows results for our hallucination network after the several variations of learning processes, different losses and with and without cross-stream connections.

Finally, the last row, #15, reports results after the last fine-tuning step which further narrows the gap with the upper bound.

Contribution of the Cross-Stream Connections. We claim that the signal injection provided by the cross-stream connections helps the learning of a better hallucination network. Row #13 and #14 show the performances for the hallucination network learning process, starting from the same point and using the same loss. The hallucination network that is learned using multiplicative connections performs better than its counterpart, where depth and RGB frames are properly aligned. It is important to note though that this is not observed in the other two smaller datasets, due to the spatial misalignment of modalities, and consequently between feature maps.

Contributions of the Proposed Distillation Loss (Eq. 5). The distillation and Euclidean losses have complementary contributions to the learning of the hallucination network. This is observed by looking at the performances reported in rows #3, #4 and #5, and also #11, #12 and #13. In both the training procedure proposed by Hoffman *et al.* [11] and our staged training process, the distillation loss improves over the Euclidean loss, and the combination of both improves over the rest. This suggests that both Euclidean and distillation losses have its own share and act differently to align the hallucination (student) and depth (teacher) feature maps and outputs' distributions.

Contributions of the Proposed Training Procedure. The intuition behind the staged training procedure proposed in this work can be ascribed to the *divide et impera* (divide-and-conquer) strategy. In our case, it means breaking the problem in two parts: learning the actual task we aim to solve and learning the student network to face test-time limitations. Row #5 reports accuracy for the architecture proposed by Hoffman *et al.*, and rows #15 report the performance for our model with connections. Both use the same loss to learn the hallucination network, and both start from the same initialization. We observe that our method outperform the one in row #5, which justifies the proposed staged training procedure.

4.3   Inference with Noisy Depth

Suppose that in a real test scenario we can only access unreliable sensors which produce noisy depth data. The question we now address is: to which extent can

we trust such noisy data? In other words, at which level of noise does it become favorable to hallucinate the depth modality with respect to using the full teacher model (step 2) with noisy depth data?

Table 2. Accuracy of the model tested with clean RGB and noisy depth data. Accuracy of the proposed hallucination model, i.e. with *no depth* at test time, is 77.21%.

σ^2	*no noise*	10^{-3}	10^{-2}	10^{-1}	10^0	10^1	*void*
Accuracy	81.43%	81.34%	81.12%	76.85%	62.47%	51.43%	14.24%

The depth sensor used in the NTU dataset (Kinect), is an IR emitter coupled with an IR camera, and has very complex noise characterization comprising at least 6 different sources [18]. It is beyond the scope of this work to investigate noise models affecting the depth channel, so, for our analysis we choose the most common one, i.e., the multiplicative speckle noise. Hence, we inject Gaussian noise in the depth images I in order to simulate speckle noise: $I = I * n, n \sim \mathcal{N}(1, \sigma)$. Table 2 shows how performances of the network degrade when depth is corrupted with such Gaussian noise with increasing variance (NTU cross-view protocol only). Results show that accuracy significantly decreases w.r.t. the one guaranteed by our hallucination model (77.21% - row #15 in Table 1), even with low noise variance. This means, in conclusion, that *training an hallucination network is an effective way not only to obviate to the problem of a missing modality, but also to deal with noise affecting the input data channel.*

4.4   Comparison with Other Methods

Table 3 compares performances of different methods on the various datasets. The standard performance measure used for this task and datasets is classification accuracy, estimated according to the protocols (training and testing splits) reported in the respective works we are comparing with.

The first part of the table (indicated by × symbol) refers to unsupervised methods, which achieve surprisingly high results even without relying on labels in learning representations.

The second part refers to supervised methods (indicated by △), divided according to the modalities used for training and testing. Here, we list the performance of the separate RGB and depth streams trained in step 1, as a reference. We expect our final model to perform better than the one trained on RGB only, whose accuracy constitutes a lower bound for our student network. The values reported for *our step 1* models for UWA3DII and NW-UCLA datasets refer to the fine-tuning of our NTU model. We have experimented training using pre-trained ImageNet weights, which led from 20% to 30% less accuracy. We also

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.

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**%	

propose our baseline, consisting in the teacher model trained in step 2. Its accuracy represents an upper bound for the final model, which will not rely on depth data at test time.

The last part of the table (indicated by □) reports our model's performances at 2 different stages together with the other privileged information method [11]. For all datasets and protocols, we can see that our privileged information approach outperforms [11], which is the only fair *direct* comparison we can make (same training & test data). Besides, as expected, our final model performs better than "Ours - RGB model, step 1" since it exploits more data at training time, and worse than "Ours - step 2", since it exploits less data at test time. Other RGB+D methods perform better (which is comprehensible since they rely on RGB+D in both training and test) but not by a large margin.

4.5 Inverting Modalities - RGB Distillation

The results presented in Table 4 address the opposite case of what is studied in the rest of the paper, *i.e.*, the case when RGB data is missing. In this case, the hallucination stream distills knowledge from the RGB stream in step 3 (Fig. 1).

We observe that the performance of the final model degrades by almost 1%, 76.41% vs. 77.21% (cf. line 15 of Table 2 in the paper). A more consistent setting would be to modify the model, inverting the cross-stream connections in Step 3 and 4, thus having information flowing again from depth to RGB.

Table 4. RGB distillation (NTU RGB-D, cross-view protocol.)

#	Method	Test modality	Loss	Cross-View
13a	Ours - step 3	Depth (*hall*)	Equation 5	76.12%
15a	Ours - step 4	Depth	x-entr	76.41%

5 Conclusions and Future Work

In this paper, we address the task of video action recognition in the context of privileged information. We propose a new learning paradigm to teach an hallucination network to mimic the depth stream. Our model outperforms many of the supervised methods recently evaluated on the NTU RGB+D dataset, as well as the hallucination model proposed in [11]. We conducted an extensive ablation study to verify how the several parts composing our learning paradigm contribute to the model performance. As a future work, we would like to extend this approach to deal with additional modalities that may be available at training time, such as skeleton joints data or infrared sequences. Finally, the current model cannot be applied to still images due to the presence of temporal convolutions. In principle, we could remove them and apply our method to still images and other tasks such as object detection.

References

1. Ba, L.J., Caruana, R.: Do deep nets really need to be deep? In: Proceedings of Advances in Neural Information Processing Systems (NIPS) (2014)
2. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 886–893. IEEE (2005)
4. Eitel, A., Springenberg, J.T., Spinello, L., Riedmiller, M., Burgard, W.: Multimodal deep learning for robust RGB-D object recognition. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 681–687. IEEE (2015)

5. Feichtenhofer, C., Pinz, A., Wildes, R.P.: Spatiotemporal multiplier networks for video action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4768–4777 (2017)
6. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1933–1941 (2016)
7. Gkioxari, G., Malik, J.: Finding action tubes. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 759–768. IEEE (2015)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
9. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 630–645. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_38
10. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: Deep Learning and Representation Learning Workshop: NIPS 2014 (2014)
11. Hoffman, J., Gupta, S., Darrell, T.: Learning with side information through modality hallucination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 826–834 (2016)
12. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1725–1732 (2014)
13. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–8. IEEE (2008)
14. Liu, J., Akhtar, N., Mian, A.: Viewpoint invariant action recognition using RGB-D videos. arXiv preprint arXiv:1709.05087 (2017)
15. Lopez-Paz, D., Bottou, L., Schölkopf, B., Vapnik, V.: Unifying distillation and privileged information. In: Proceedings of the International Conference on Learning Representations (ICLR) (2016)
16. Luo, Z., Jiang, L., Hsieh, J.T., Niebles, J.C., Fei-Fei, L.: Graph distillation for action detection with privileged information. arXiv preprint arXiv:1712.00108 (2017)
17. Luo, Z., Peng, B., Huang, D.A., Alahi, A., Fei-Fei, L.: Unsupervised learning of long-term motion dynamics for videos. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). No. EPFL-CONF-230240 (2017)
18. Mallick, T., Das, P.P., Majumdar, A.K.: Characterizations of noise in kinect depth images: a review. IEEE Sens. J. **14**(6), 1731–1740 (2014). https://doi.org/10.1109/JSEN.2014.2309987
19. Ohn-Bar, E., Trivedi, M.M.: Joint angles similarities and HOG2 for action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 465–470. IEEE (2013)
20. Rahmani, H., Bennamoun, M.: Learning action recognition model from depth and skeleton videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5832–5841 (2017)
21. Rahmani, H., Mahmood, A., Huynh, D., Mian, A.: Histogram of oriented principal components for cross-view action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **38**(12), 2430–2443 (2016)

22. Rahmani, H., Mian, A., Shah, M.: Learning a deep model for human action recognition from novel viewpoints. IEEE Trans. Pattern Anal. Mach. Intell. **40**(3), 667–681 (2018)
23. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: NTU RGB+D: A large scale dataset for 3D human activity analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1010–1019 (2016)
24. Shahroudy, A., Ng, T.T., Gong, Y., Wang, G.: Deep multimodal feature analysis for action recognition in RGB+D videos. IEEE Trans. Pattern Anal. Mach. Intell. **40**(5), 1045–1058 (2018)
25. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in Neural Information Processing Systems, pp. 568–576 (2014)
26. Soo Kim, T., Reiter, A.: Interpretable 3D human action analysis with temporal convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 20–28 (2017)
27. Sun, L., Jia, K., Yeung, D.Y., Shi, B.E.: Human action recognition using factorized spatio-temporal convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4597–4605 (2015)
28. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4489–4497 (2015)
29. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6450–6459 (2018)
30. Vapnik, V., Vashist, A.: A new learning paradigm: learning using privileged information. Neural Netw. **22**(5), 544–557 (2009)
31. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3169–3176. IEEE (2011)
32. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3551–3558 (2013)
33. Wang, J., Nie, X., Xia, Y., Wu, Y., Zhu, S.C.: Cross-view action modeling, learning and recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2649–2656 (2014)
34. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. arXiv preprint arXiv:1711.07971 (2017)