Learning with Privileged Information via Adversarial Discriminative Modality Distillation

通過對抗性判別模態蒸餾(Adversarial Discriminative Modality Distillation) 學習特權資訊 (Privileged Information)

Nuno C. Garcia, Pietro Morerio, and Vittorio Murino, Senior Member, IEEE

https://ieeexplore.ieee.org/document/8764498

Comments: Accepted to IEEE Transactions on Pattern Analysis and Machine Intelligence
Subjects: Computer Vision and Pattern Recognition (cs.CV)

Abstract 摘要

Heterogeneous data modalities can provide complementary cues for several tasks, usually leading to more robust algorithms and better performance.
異構資料模式可以為多個任務提供補充線索，通常會導致更強大的演算法和更好的性能。

However, while training data can be accurately collected to include a variety of sensory modalities, it is often the case that not all of them are available in real life (testing) scenarios, where a model has to be deployed.
然而，雖然可以準確地收集訓練資料以包括各種感官模式，但在現實生活（測試）場景中，通常情況下並非所有這些都可用，在這些場景中必須部署模型。

This raises the challenge of how to extract information from multimodal data in the training stage, in a form that can be exploited at test time, considering limitations such as noisy or missing modalities.
這對來自訓練階段的多模態資料中如何提出訊息提出了挑戰，以一種可以在測試時利用的形式，考慮到諸如嘈雜或缺失模態等限制。

This paper presents a new approach in this direction for RGB-D vision tasks, developed within the adversarial learning and privileged information frameworks.
本文提出了一種針對 RGB-D 視覺任務的新方法，該方法是在對抗性學習和特權資訊框架內開發的。

We consider the practical case of learning representations from depth and RGB videos, while relying only on RGB data at test time.
我們考慮從深度和 RGB 影像中學習表示的實際案例，而在測試時僅依賴於 RGB 資料。

We propose a new approach to train a hallucination network that learns to distill depth information via adversarial learning, resulting in a clean approach without several losses to balance or hyperparameters.
我們提出了一種訓練幻覺網絡的新方法，該方法通過對抗性學習來學習提取深度資訊，從而產生一種

乾淨的方法，而不會出現一些平衡或超參數損失。

We report state-of-the-art results for object classification on the NYUD dataset, and video action recognition on the largest multimodal dataset available for this task, the NTU RGB+D, as well as on the Northwestern-UCLA.
我們報告了 NYUD 資料集上對象分類的最新結果，以及可用於此任務的最大多模式資料集 NTU RGB+D 以及 Northwestern-UCLA 上的影像動作識別結果。

Index Terms—Multimodal deep learning, adversarial learning, privileged information, network distillation, modality hallucination.

1 INTRODUCTION 前言

DEPTH perception is the ability to reason about the 3D world, critical for the survival of many hunting predators and an important skill for humans to understand and interact with the surrounding environment.
深度感知是對 3D 世界進行推理的能力，對許多捕食者的生存至關重要，也是人類理解周圍環境並與之互動的重要技能。

It develops very early in humans when babies start to crawl [1], and emerges from a variety of mechanisms that jointly contribute to the sense of relative and absolute position of objects, called depth cues.
當嬰兒開始爬行時，它在人類的早期發育 [1]，並從各種機制中出現，這些機制共同促成物體的相對和絕對位置感，稱為深度線索。

Besides binocular cues (e.g. stereovision), humans use monocular cues that relate to a priori visual assumptions derived from 2D single images through shadows, perspective, texture gradient, and other signals (e.g. the assumption that objects look blurrier the further they are, or that if an object occludes another it must be closer, etc.) [2].
除了雙目線索（例如立體視覺），人類還使用與通過陰影、透視、紋理梯度和其他信號（例如，物體越遠看起來越模糊的假設，或 如果一個物體遮擋了另一個物體，它必須更近，等等）[2]。

As matter of fact, although humans underestimate object distance in a monocular vision setup [3], we are still able to perform most of our vision-related tasks with good efficiency even with one eye covered.
事實上，儘管人類在單目視覺設置中低估了物距 [3]，但即使遮住一隻眼睛，我們仍然能夠以良好的效率執行大部分與視覺相關的任務。

Similarly, depth perception is often of paramount importance for many computer vision tasks related to robotics, autonomous driving, scene understanding, to name a few.
同樣，對於與機器人、自動駕駛、場景理解等相關的許多計算機視覺任務，深度感知通常是最重要的。

The emergence of cheap depth sensors and the need for big data led to big multimodal datasets containing RGB, depth, infrared, and skeleton sequences [4], which in turn stimulated multimodal deep learning approaches.
廉價深度傳感器的出現和對大數據的需求導致了包含 RGB、深度、紅外和骨架序列的大型多模態資料集 [4]，進而刺激了多模態深度學習方法。

Traditional computer vision tasks like action recognition, object detection, or instance segmentation have been shown to benefit performance gains if the model considers other modalities, namely depth, instead of RGB only [5], [6], [7], [8].
如果模型考慮其他模式，即深度，而不是僅考慮 RGB [5]、[6]、[7]、[8]，則傳統的計算機視覺任務（如動作識別、對象檢測或實例分割）已被證明有助於提高性能 .

However, it is reasonable to expect that depth data is not going to be always available when a model is deployed in real scenarios, either due to the impossibility to collect depth data with enough quality (e.g., due to far-distance or reflectance issues) or to install depth sensors everywhere, sensor or communications failure, or other unpredictable events.
然而，在實際場景中部署模型時，由於無法收集足夠質量的深度資料（例如，由於遠距離或反射問題），預計深度資料並不總是可用是合理的。 或到處安裝深度傳感器，傳感器或通信故障，或其他不可預測的事件。

Considering this limitation, we would like to answer the following question (also depicted in Fig. 1): what is the best way of using all data available at training time, in order to learn robust representations, knowing that there are missing (or noisy) modalities at test time?
考慮到這個限制，我們想回答以下問題（也在圖 1 中描繪）：在訓練時使用所有可用資料的最佳方法是什麼，以便學習穩健的表示，知道存在缺失（或噪聲）) 測試時的模式？

In other words, is there any added value in training a model by exploiting multimodal data, even if only one modality is available at test time?
換句話說，即使在測試時只有一種模態可用，通過利用多模態資料來訓練模型是否有任何附加價值？

Unsurprisingly, the simplest and most commonly adopted solution consists in training the model using only the modality in which it will be tested.
不出所料，最簡單和最常用的解決方案是僅使用將要測試的模態來訓練模型。

Nevertheless, a more interesting alternative is to exploit the potential of the available data and train the model using all modalities, being however aware of the fact that not all of them will be accessible at test time.
然而，更有趣的替代方法是利用可用資料的潛力並使用所有模式訓練模型，但要意識到在測試時並非所有模式都可以訪問的事實。

This learning paradigm, i.e., when the model is trained using extra information, is generally known as learning with privileged information [9] or learning with side information [10].

這種學習範式，即當模型使用額外資訊進行訓練時，通常被稱為使用特權資訊學習 [9] 或使用輔助資訊學習 [10]。

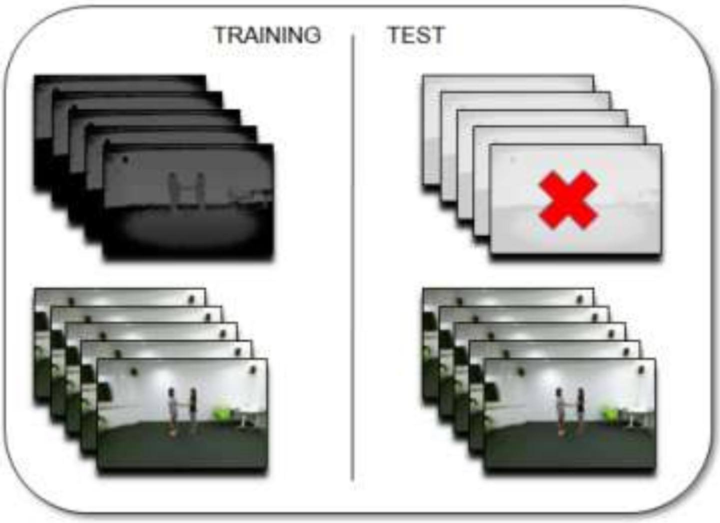


Fig. 1. What is the best way of using all data available at training time, considering a missing (or noisy) modality at test time?

Fig. 1. What is the best way of using all data available at training time, considering a missing (or noisy) modality at test time?

圖 1. 考慮到測試時缺失（或嘈雜）的模態，在訓練時使用所有可用資料的最佳方法是什麼？

In this work, we propose an adversarial discriminative modality distillation (ADMD) strategy within a multimodalstream framework that learns from different data modalities and can be deployed and tested on a subset of these.

在這項工作中，我們在多模態流框架內提出了一種對抗性判別模態蒸餾 (ADMD) 策略，該策略從不同的資料模態中學習，並且可以在其中的一個子集上進行部署和測試。

Particularly, our model learns from RGB and depth video sequences and is tested on RGB only.

特別是，我們的模型從 RGB 和深度影像序列中學習，並且僅在 RGB 上進行了測試。

Still, due to its general design it can also be used with whatever combination of other modalities as well.

儘管如此，由於其通用設計，它還可以與其他模式的任何組合一起使用。

We evaluate its performance on the task of video action recognition and object classification.

我們評估其在影像動作識別和對象分類任務上的表現。

To this end, we introduce a new adversarial learning strategy to learn a hallucination network (Fig. 2), whose goal is to mimic the test time missing modality features, while preserving their discriminative power.

為此，我們引入了一種新的對抗性學習策略來學習幻覺網絡（圖 2），其目標是模仿測試時間缺失的模態特徵，同時保留其判別力。

The hallucination network uses RGB only as input and tries to recover useful depth features for the task at hand.
幻覺網絡僅使用 RGB 作為輸入，並嘗試為手頭的任務恢復有用的深度特徵。

Such network can be thought as a source of the aforementioned monocular depth cues, i.e., a source of depth cues from a single 2D RGB image.
這種網絡可以被認為是上述單眼深度線索的來源，即來自單個 2D RGB 圖像的深度線索的來源。

We would like to stress the fact that, in contrast to estimating real depth maps from RGB, we operate at feature level.
我們想強調一個事實，與從 RGB 估計真實深度圖相比，我們在特徵級別進行操作。

Conceptually, it may seem that directly estimating depth maps from RGB is a more straightforward approach to deal with missing depth at test time.
從概念上講，直接從 RGB 估計深度圖似乎是在測試時處理缺失深度的更直接的方法。

However, this is arguably a much more difficult task to accomplish compared to the primary task at hand, which is action/object recognition from RGB sequences.
然而，與手頭的主要任務（從 RGB 序列識別動作/對象）相比，這可以說是一項更難完成的任務。

A more reasonable approach is to reduce the depth estimation problem from the pixel space to a low dimensional space, while continuing to profit to some extent of the discriminative benefits offered by the depth modality.
更合理的方法是將深度估計問題從像素空間減少到低維空間，同時繼續在一定程度上受益於深度模態提供的判別優勢。

On the one hand, our work is inspired by previous works using hallucination networks in the context of learning with privileged information.
一方面，我們的工作受到以前在特權資訊學習背景下使用幻覺網絡的工作的啟發。

This was primarily proposed in [10], that presented an end-to-end single step training method to learn a hallucination network.
這主要是在 [10] 中提出的，它提出了一種端到端的單步訓練方法來學習幻覺網絡。

This work was recently revisited in [11] considering a multi-step learning paradigm using a loss inspired by the generalized distillation framework [12].
最近在 [11] 中重新審視了這項工作，考慮到使用受廣義蒸餾框架 [12] 啟發的損失的多步學習範式。

On the other hand, adversarial learning has been shown to be a powerful tool to model data distributions [13], [14].

另一方面，對抗性學習已被證明是對資料分佈進行建模的強大工具 [13]、[14]。

Building upon these ideas, we propose a new approach to learn the hallucination network via a discriminative adversarial learning strategy.
基於這些想法，我們提出了一種通過區分性對抗學習策略來學習幻覺網絡的新方法。

Our proposed method has several advantages: it is agnostic regarding the pair of modalities used, which greatly simplifies its extension beyond RGB and depth data; and it is able to deal with videos by design, by exploiting a form of temporal supervision as auxiliary information.
我們提出的方法有幾個優點：它與所使用的模態對無關，這大大簡化了它在 RGB 和深度資料之外的擴展； 它能夠通過設計來處理影像，利用一種形式的時間監督作為輔助資訊。

Furthermore, it dumps the need to balance the different losses used in the other methods [10] [11].
此外，它不需要平衡其他方法中使用的不同損失 [10] [11]。

Finally, thanks to the discriminator design, which includes an auxiliary classification task, our method is able to transfer the discriminative capability from a so-called teacher network [12] (depth network) to a student (hallucination network), up to a full recovery of the teacher's accuracy.
最後，由於鑑別器設計，其中包括一個輔助分類任務，我們的方法能夠將鑑別能力從所謂的教師網絡 [12]（深度網絡）轉移到學生（幻覺網絡），達到一個完整的 恢復老師的準確性。

To summarize, the main contributions of this paper are the following:
總而言之，本文的主要貢獻如下：

We propose a new approach to learn a hallucination network within a multimodal-stream network architecture:
我們提出了一種在多模式流網絡架構中學習幻覺網絡的新方法：

it consists in an adversarial learning strategy that exploits multiple data modalities at training while using only one at test time.
它包含一種對抗性學習策略，該策略在訓練時利用多種資料模態，而在測試時僅使用一種。

It proved to outperform its distance-based method counterparts [10], [11], and to augment its flexibility by being agnostic to components like distance metrics, data modalities, and size of the hallucinated feature vectors.
事實證明，它優於基於距離的方法[10]、[11]，並且通過對距離度量、資料模態和幻覺特徵向量的大小等組件不可知來增強其靈活性。

More technically, we propose a discriminator network which is time-aware, and jointly solves 1) the classical binary classification task (real/generated), and 2) an auxiliary task, which inherently endows the learned features with discriminative power.

從技術上講，我們提出了一個時間感知的鑑別器網絡，並聯合解決了 1）經典的二元分類任務（真實/生成），以及 2）一個輔助任務，它固有地賦予了學習特徵以鑑別能力。

We report results – in the privileged information scenario – on the NYUD [15] dataset for the task of object classification, and on the large-scale NTU RGB+D [16] and the Northwestern-UCLA [17] datasets for the task of action recognition.
我們在特權資訊場景中報告了用於對象分類任務的 NYUD [15] 資料集，以及用於執行以下任務的大規模 NTU RGB+D [16] 和 Northwestern-UCLA [17] 資料集的結果 動作識別。

The rest of the paper is organized as follows.
本文的其餘部分安排如下。

Section 2 relates this work to the literature in privileged information, multimodal deep learning, and adversarial learning.
第 2 節將這項工作與特權資訊、多模態深度學習和對抗性學習方面的文獻聯繫起來。

Section 3 presents the details of the proposed architecture and the novel learning strategy.
第 3 節介紹了所提出的架構和新穎的學習策略的細節。

Section 4 reports results on object recognition and video action recognition datasets, comparing them to the current state of the art, and investigating how the different parts of our approach contribute to the overall performance through an extensive ablation study.
第 4 節報告了對象識別和影像動作識別資料集的結果，將它們與當前最先進的技術進行比較，並通過廣泛的消融研究調查我們方法的不同部分如何對整體性能做出貢獻。

Finally, we draw conclusions and future research directions in Section 5.
最後，我們在第 5 節中得出結論和未來的研究方向。

2 RELATED WORK 相關工作

Our work is at the intersection of four topics: adversarial learning [13], RGB-D vision, network distillation [18] and privileged information [9].
我們的工作處於四個主題的交叉點：對抗性學習 [13]、RGB-D 視覺、網絡蒸餾 [18] 和特權資訊 [9]。

As Lopez et al. noted, privileged information and network distillation are instances of the same more inclusive theory, called generalized distillation [12].
正如 Lopez et al。 注意到，特權資訊和網絡蒸餾是相同的更具包容性的理論的實例，稱為廣義蒸餾 [12]。

2.1 Generalized Distillation 廣義蒸餾

Within the generalized distillation framework, our model is both related to the privileged information theory [9], considering that the extra modality (depth, in this case) is only used at training time; and to the distillation framework, considering that our hallucination network is effectively learning by distilling the knowledge of a the previously learned ”teacher” network, despite not using a distillation loss.
在廣義蒸餾框架內，我們的模型都與特權資訊理論 [9] 相關，考慮到額外的模態（在這種情況下為深度）僅在訓練時使用； 以及蒸餾框架，考慮到我們的幻覺網絡通過蒸餾先前學習的“教師”網絡的知識而有效地學習，儘管沒有使用蒸餾損失。

In this context, the closest works to our approach are [10] by Hoffman et al. and [11] by Garcia et al.
在這種情況下，與我們的方法最接近的作品是 Hoffman et al [10]。 和 [11] 由 Garcia et al. 撰寫。

The work of Hoffman et al. [10] introduced a model to hallucinate depth features from RGB input for object detection task.
Hoffman et al. [10]的工作，引入了一個模型，用於從 RGB 輸入中產生幻覺深度特徵，用於對象檢測任務。

While the idea of using a hallucination stream is similar to the one thereby presented, the mechanism used to learn it is different.
雖然使用幻覺流的想法與由此提出的想法相似，但用於學習它的機制是不同的。

In [10], the authors use an Euclidean loss between the depth and hallucinated feature maps, that is part of the total loss along with more than ten classification and localization losses.
在 [10] 中，作者在深度和幻覺特徵圖之間使用了歐幾里得損失，這是總損失的一部分以及十多個分類和定位損失。

This makes its effectiveness dependent on hyperparameter tuning to balance the different values, as the model is trained jointly in one step by optimizing the aforementioned composite loss.
這使得它的有效性取決於超參數調整以平衡不同的值，因為模型是通過優化上述複合損失在一步中聯合訓練的。

In [11], Garcia et al. built on this idea to propose a new staged training procedure that lead to learn a better teacher network, and a new loss inspired from the distillation framework.
在 [11] 中，Garcia et al. 基於這個想法，提出了一個新的分階段訓練程序，可以學習更好的教師網絡，以及從蒸餾框架中獲得靈感的新損失。

This loss is composed by the Euclidean distance between feature maps, the cross-entropy using the ground truth labels, and a cross-entropy using as targets the soft predictions from the teacher network (the depth stream, in this case).
該損失由特徵圖之間的歐幾里得距離、使用真實標籤的交叉熵以及使用教師網絡（在這種情況下為深度流）的軟預測作為目標的交叉熵組成。

Moreover, the authors encouraged the learning by design, by using multiplier cross-stream connections [19], and extended the method to video action recognition.
此外，作者鼓勵通過設計學習，通過使用乘數交叉流連接 [19]，並將方法擴展到影像動作識別。

Differently from [11] and [10], we propose an adversarial strategy to learn the hallucination stream, which alleviates the need for balancing losses or tuning hyperparameters.
與 [11] 和 [10] 不同，我們提出了一種對抗性策略來學習幻覺流，這減輕了平衡損失或調整超參數的需要。

An interesting work lying at the intersection of multimodal learning and learning with privileged information is ModDrop by Neverova et al. [20].
Neverova et al. 的 ModDrop 是多模態學習和特權資訊學習交叉點的一項有趣工作[20]。

Here the authors propose a modality-based dropout strategy, where each input modality is entirely dropped (actually zeroed) with some probability during training.
在這裡，作者提出了一種基於模態的 dropout 策略，其中每個輸入模態在訓練期間以一定的概率被完全丟棄（實際上為零）。

The resulting model is proved to be more resilient to missing modalities at test time.
結果模型被證明對測試時丟失的模態更有彈性。

We compare with ModDrop in the task of object classification. Luo et al. [21] addressed a similar problem, where the model is first trained on several modalities (RGB, depth, optical flow, and joints), but tested only in one.
我們在對象分類的任務中與 ModDrop 進行了比較。 羅等人。 [21] 解決了一個類似的問題，該模型首先在幾種模式（RGB、深度、光流和關節）上進行訓練，但僅在一種模式下進行了測試。

The authors propose a graph-based distillation method to distill information across all modalities at training time, allowing each modality to learn from all others.
作者提出了一種基於圖的蒸餾方法，可以在訓練時從所有模態中提取資訊，允許每種模態向所有其他模態學習。

This approach achieves state-of-the-art results in action recognition and action detection tasks.
這種方法在動作識別和動作檢測任務中取得了最先進的結果。

Our work substantially differs from [21] since we benefit from a hallucination mechanism, consisting in an auxiliary hallucination network trained by leveraging a previously trained network.
我們的工作與 [21] 有很大不同，因為我們受益於幻覺機制，包括通過利用先前訓練的網絡訓練的輔助幻覺網絡。

This mechanism allows the model to learn to emulate the presence of the missing modality at test time.

這種機制允許模型學習在測試時模擬缺失模態的存在。

Another recent related work is [22], which proposes a distillation framework where there is no frozen teacher network, but all the networks work as an ensemble that learn in a collaboratively manner.
最近的另一項相關工作是 [22]，它提出了一個蒸餾框架，其中沒有凍結的教師網絡，但所有網絡都作為一個以協作方式學習的整體工作。

Learning with privileged information for action recognition has also been explored for recurrent neural networks, e.g. in [23] where the authors devise a method that uses skeleton joints as privileged information to learn a better action classifier that uses depth, even with scarce data.
還探索了使用用於動作識別的特權資訊進行學習的循環神經網絡，例如 在 [23] 中，作者設計了一種方法，該方法使用骨骼關節作為特權資訊來學習使用深度的更好的動作分類器，即使資料稀少。

2.2 RGB-D vision RGB-D 視覺

Video action recognition and object detection have a long and rich field of literature, spanning from classification methods using handcrafted features, e.g. [24], [25], [26], [27], [28], [29] to modern deep learning approaches, e.g. [7], [30], [31], [32], [33], using either RGB-only or together with depth data.
影像動作識別和對象檢測擁有悠久而豐富的文獻領域，涵蓋使用手工特徵的分類方法，例如 [24]、[25]、[26]、[27]、[28]、[29] 到現代深度學習方法，例如 [7]、[30]、[31]、[32]、[33]，僅使用 RGB 或與深度資料一起使用。

We point to some of the more relevant works in video action recognition and object recognition using RGB and depth, including state-of-the-art methods considering the NTU RGB+D and the NYU-Depth V2 datasets, as well as architectures related to our proposed model.
我們指出了使用 RGB 和深度的影像動作識別和對象識別中一些更相關的工作，包括考慮 NTU RGB+D 和 NYU-Depth V2 資料集的最先進方法，以及與 我們提出的模型。

2.2.1 Video action recognition 影像動作識別

The two-stream model introduced by Simonyan and Zisserman [34] is a landmark on video analysis, and since then has inspired a series of variants that achieved state-of-theart performance on diverse datasets.
Simonyan 和 Zisserman [34] 引入的雙流模型是影像分析的里程碑，從那時起激發了一系列變體，在不同的資料集上實現了最先進的性能。

This architecture is composed by a RGB and an optical flow stream, which are trained separately, and then fused at the prediction layer.
該架構由 RGB 和光流流組成，分別訓練，然後在預測層融合。

Our model relates to this since the test-time predictions result from the average of the hallucination stream and the RGB stream logits.

我們的模型與此相關，因為測試時間預測來自幻覺流 (the hallucination stream) 和 RGB 流 logits (RGB stream) 的平均值。

In [19], the authors propose a variation of the latter, which models spatiotemporal features by injecting the motion stream's signal into the residual unit of the appearance stream.
在[19]中，作者提出了後者的一種變體，它通過將運動流的信號注入外觀流的殘差單元來對時空特徵進行建模。

They also employ 1D temporal convolutions along with 2D spatial convolutions, which we also adopt in this paper.
他們還使用了 1D 時間卷積和 2D 空間卷積，我們在本文中也採用了這種卷積。

Indeed, the combination of 2D spatial and 1D temporal convolutions has shown to learn better spatiotemporal features than 3D convolutions [35].
事實上，2D 空間和 1D 時間卷積的組合已表明比 3D 卷積學習更好的時空特徵 [35]。

The current state of the art in video action recognition [36] uses 3D temporal convolutions and a new building block dedicated to capture long range dependencies, using RGB data only.
當前影像動作識別領域的最新技術 [36] 使用 3D 時間卷積和一個新的構建塊，專門用於捕獲長距離依賴關係，僅使用 RGB 資料。

Instead, in [5] the authors explore the complementary properties of RGB and depth data, taking the NTU RGB+D dataset as testbed.
相反，在 [5] 中，作者探索了 RGB 和深度資料的互補特性，將 NTU RGB+D 資料集作為測試平台。

They propose a deep autoencoder architecture and a structured sparsity learning machine, and achieve state-of-the-art results for action recognition.
他們提出了深度自動編碼器架構和結構化稀疏學習機，並在動作識別方面取得了最先進的結果。

Liu et al. [6] also use RGB and depth to devise a method for viewpoint invariant action recognition.
Liu et al. [6] 還使用 RGB 和深度設計了一種用於視點不變動作識別的方法。

First, the method extracts dense trajectories from RGB data, which are then encoded in viewpoint invariant deep features.
首先，該方法從 RGB 資料中提取密集軌跡，然後將其編碼為視點不變的深度特徵。

The RGB and depth features are then used as a dictionary for test time prediction.
然後將 RGB 和深度特徵用作測試時間預測的字典。

To the best of our knowledge, these are state-of-the-art approaches to exploit RGB+D for video action recognition, that report results on the NTU RGB+D dataset [16], the largest video action recognition dataset to

offer RGB and depth.
據我們所知，這些是利用 RGB+D 進行影像動作識別的最先進方法，在 NTU RGB+D 資料集 [16] 上報告結果，這是提供 RGB 的最大影像動作識別資料集 和深度。

It is important to note that we propose a fully convolutional model that exploits RGB and depth data at training time only, and uses exclusively RGB data as input at test time.
需要注意的是，我們提出了一個完全卷積模型，該模型僅在訓練時利用 RGB 和深度資料，並在測試時僅使用 RGB 資料作為輸入。

This work goes in the direction of reducing the performance gap between privileged information and traditional approaches.
這項工作的方向是縮小特權資訊與傳統方法之間的性能差距。

2.2.2 Object recognition 物體識別

Over the years, object recognition based on RGB and depth have been an insightful task to reason on the complementarity of these two modalities, and whether depth data should be handled differently, compared to RGB.
多年來，基於 RGB 和深度的對象識別一直是推理這兩種模式的互補性以及與 RGB 相比是否應該以不同方式處理深度資料的有見地的任務。

An example of this is [7], in which the authors propose to encode depth images using a geocentric embedding that encodes height above ground and angle with gravity for each pixel in addition to the horizontal disparity, showing that it works better than using raw depth.
這方面的一個例子是 [7]，其中作者建議使用地心嵌入對深度圖像進行編碼，該嵌入對每個像素的地面高度和重力角度以及水平視差進行編碼，表明它比使用原始深度效果更好 .

Differently, in [33], the authors focus on carefully designing a convolutional neural network including a multimodal layer to fuse RGB and depth.
不同的是，在 [33] 中，作者專注於精心設計一個卷積神經網絡，包括一個多模態層來融合 RGB 和深度。

Our work differs from these approaches since we focus on learning a model that has access to depth only at training time, which fundamentally changes the feature learning approach.
我們的工作與這些方法不同，因為我們專注於學習只能在訓練時訪問深度的模型，這從根本上改變了特徵學習方法。

2.3 Adversarial Learning 對抗學習

In the seminal paper of Goodfellow et al. [13], the authors propose a generative model that is trained by having two networks playing the so called minimax game.

在 Goodfellow et al. [13] 的開創性論文中，作者提出了一個生成模型，該模型通過讓兩個網絡玩所謂的極小極大遊戲來訓練。

A generator network is trained to generate images from noise vectors, and a discriminator network is trained to classify the generated images as false, and images sampled from the dataset as true.
訓練生成器網絡從噪聲向量生成圖像，訓練鑑別器網絡將生成的圖像分類為假，並將從資料集中採樣的圖像分類為真。

As the game evolves, the generator becomes better and better at generating samples that look like the true images from the data distribution.
隨著遊戲的發展，生成器越來越擅長從資料分佈中生成看起來像真實圖像的樣本。

Many papers extended this approach in different directions, such as disentangling semantic concepts [37], network compression [38] [39] [40], feature augmentation [41], image to image translation [42], and explored different losses [43] and other tricks to improve performance and stability [44] [45].
許多論文在不同的方向上擴展了這種方法，例如解開語義概念 [37]、網絡壓縮 [38] [39] [40]、特徵增強 [41]、圖像到圖像翻譯 [42]，並探索了不同的損失 [43]] 和其他提高性能和穩定性的技巧 [44] [45]。

Our work relates to this body of work, as the hallucination network of our model tries to generate features from the missing modality feature space through adversarial learning.
我們的工作與這一系列工作有關，因為我們模型的幻覺網絡試圖通過對抗性學習從缺失的模態特徵空間中生成特徵。

However, not only the context here is different, since adversarial learning is explored in the framework of privileged information, but also the task assigned to the discriminator is not the one typically used in adversarial learning, as detailed in Section 3.
然而，不僅這裡的上下文不同，因為對抗性學習是在特權資訊的框架中探索的，而且分配給鑑別器的任務也不是對抗性學習中通常使用的任務，如第 3 節所述。

An important variant of the GAN framework are Conditional GANs (CGANs) [46], that propose to concatenate the label of desired class to be generated, to the noise vector.
GAN 框架的一個重要變體是條件 GAN（CGAN）[46]，它建議將要生成的所需類別的標籤連接到噪聲向量。

This mechanism is related, yet different, to how our generator network is implicitly conditioned in this work.
這種機制與我們的生成器網絡在這項工作中的隱式條件有關，但又有所不同。

Our generator network input is a small volume of 5 RGB frames, and temporal convolutions are zero-padded to maintainthe volume’s size along the time dimension.
我們的生成器網絡輸入是 5 個 RGB 幀的小體積，並且時間卷積被零填充以保持體積沿時間維度的大

小。

The generator is thus implicitly aware of the temporal ordering, since features generated for the first and last frames will heavily be affected by the border effect of zero padding, which is performed several times (at each residual block).
因此，生成器隱含地知道時間順序，因為為第一幀和最後一幀生成的特徵將嚴重受到零填充的邊界效應的影響，零填充會執行多次（在每個殘差塊）。

To solve this issue, we provide the temporal ordering conditioning label to the discriminator as well.
為了解決這個問題，我們還為鑑別器提供了時間排序條件標籤。

The CGAN model has been used in different domains, from image synthesis [47] to domain adaptation [41].
CGAN 模型已用於不同領域，從圖像合成 [47] 到域適應 [41]。

Perhaps more similar to our work is the recent paper by Roheda et al. [48], that also approaches the problem of missing modalities in the context of adversarial learning.
也許與我們的工作更相似的是 Roheda et al. [48] 最近的論文，這也解決了在對抗性學習中缺少模態的問題。

They address the binary task of person detection using images, seismic, and acoustic sensors, where the latter two are absent at test time.
它們使用圖像、地震和聲學傳感器解決人員檢測的二元任務，而後兩者在測試時不存在。

A CGAN is conditioned on the available images and the generator maps a vector noise to representative information from the missing modalities, with an auxiliary L2 loss.
CGAN 以可用圖像為條件，生成器將矢量噪聲映射到來自缺失模態的代表性資訊，並帶有輔助 L2 損失。

In contrast to this work, our CGAN model learns a mapping directly from the test modality to the feature space of the missing modality, with no auxiliary loss.
與這項工作相比，我們的 CGAN 模型直接從測試模態到缺失模態的特徵空間學習映射，沒有輔助損失。

Besides, we propose a two-step training procedure in order to learn a better teacher network, and provide a stable target for the generator.
此外，我們提出了一個兩步訓練程序，以學習更好的教師網絡，並為生成器提供一個穩定的目標。

Finally, we focus on the arguably more demanding tasks of video action recognition and object recognition.
最後，我們專注於影像動作識別和物體識別的要求更高的任務。

3 LEARNING TO HALLUCINATE DEPTH FEATURES 學習幻覺深度特徵

Our goal is to train a hallucination network that, having as input RGB, is able to produce similar features to the ones produced by the depth network.
我們的目標是訓練一個幻覺網絡，該網絡以 RGB 為輸入，能夠產生與深度網絡產生的特徵相似的特徵。

The reasoning behind this idea is that on one hand depth and RGB provide complementary information for the task, but on the other hand RGB alone contains some cues for depth perception.
這個想法背後的原因是，一方面深度和 RGB 為任務提供了補充資訊，但另一方面，RGB 本身就包含一些深度感知的線索。

Therefore, the goal of the hallucination network is to extract from RGB frames the complementary information that depth data would provide.
因此，幻覺網絡的目標是從 RGB 幀中提取深度資料將提供的補充資訊。

It is important to emphasize that we are interested in recovering useful depth features, in contrast to estimating real depth maps from RGB.
需要強調的是，與從 RGB 估計真實深度圖相比，我們對恢復有用的深度特徵感興趣。

This is accomplished in a two-step training procedure, illustrated in Fig. 2, and described in the following.
這是在兩步訓練過程中完成的，如圖 2 所示，並在下面描述。

The first step (Fig. 2, top) consists in training the RGB and depth streams individually, with the respective input modality, as two standard, separate, supervised learning problems.
第一步（圖 2，頂部）包括分別訓練 RGB 和深度流，使用各自的輸入模態，作為兩個標準的、獨立的、有監督的學習問題。

The resulting ensemble, obtained by fusing the predictions of the two sub-networks (not fine-tuned), represents the full model (two-stream) that can be used when both modalities are available at test time.
通過融合兩個子網絡的預測（未微調）獲得的最終集成代表了在測試時兩種模式都可用時可以使用的完整模型（雙流）。

Its accuracy should be taken as an upper bound for the model we are proposing.
它的準確性應作為我們提出的模型的上限。

In the second step (Fig. 2, bottom), we actually train the hallucination network by means of the proposed adversarial learning strategy.
在第二步（圖 2，底部）中，我們實際上是通過提出的對抗性學習策略來訓練幻覺網絡。

As the hallucination network is trained in the context of adversarial learning to generate depth features, it can be also interpreted as the generator network in the traditional GAN framework [13].

由於幻覺網絡是在對抗性學習的背景下訓練生成深度特徵的，因此也可以解釋為傳統 GAN 框架中的生成器網絡 [13]。

However, strictly speaking, it is clearly to be considered as an encoder, which tries to extract monocular depth features from RGB input data.
然而，嚴格來說，它顯然被認為是一個編碼器，它試圖從 RGB 輸入資料中提取單眼深度特徵。

The test time setup of step 2 is again a two-stream model (not fine-tuned), composed by the RGB and hallucination networks, both having RGB data as input.
步驟 2 的測試時間設置同樣是一個雙流模型（未微調），由 RGB 和幻覺網絡組成，兩者都以 RGB 資料作為輸入。

3.1 Training procedure 培訓程序

Inspired by the generalized distillation paradigm, we follow a staged learning procedure, where the “teacher” net is trained first (Step 1) and separately from the ”student” (Step 2).
受廣義蒸餾範式的啟發，我們遵循分階段學習程序，其中首先訓練“教師”網絡（步驟 1）並與“學生”網絡分開（步驟 2）。

This is in contrast with [10], where everything is learned end-to-end, but in line with [11], where separated learning steps proved to be more effective.
這與 [10] 形成對比，其中一切都是端到端學習的，但與 [11] 一致，其中分離的學習步驟被證明更有效。

Step 1. The RGB and depth streams are trained separately, which is common practice in two-stream architectures.
步驟 1. RGB 和深度流分別訓練，這是雙流架構中的常見做法。

Both depth and appearance streams are trained by minimizing the cross-entropy loss, after being initialized with a pre-trained ImageNet model for all experiments as common practice [10], [11], [21].
在使用預訓練的 ImageNet 模型初始化所有實驗後，深度和外觀流都是通過最小化交叉熵損失來訓練的，作為常見做法 [10]、[11]、[21]。

We test both streams individually and in a two-stream setup, where the final prediction results from the average of the two streams’logits.
我們分別測試兩個流並在雙流設置中測試，其中最終預測結果來自兩個流的 logits 的平均值。

We found that fine-tuning the two-stream model does not increase performance consistently.
我們發現微調雙流模型並不能始終如一地提高性能。

This step can also be regarded as training the teacher network - depth stream - for the next step (see Fig. 2,

top).
這一步也可以看作是訓練教師網絡—深度流—用於下一步（見圖 2，頂部）。

Step 2. The depth stream Ed, trained in the previous step, is now frozen, in order to provide a stable target for the hallucination network (generator) H, which plays the adversarial game with a discriminator D (see Fig. 2, bottom).
步驟 2. 上一步訓練的深度流 Ed 現在被凍結，以便為幻覺網絡（生成器）H 提供一個穩定的目標，它與判別器 D 進行對抗性遊戲（見圖 2，底部） ）。

The primary task of the discriminator D is to distinguish between the features FH generated by the hallucination network H and Fd generated by the depth network Ed.
鑑別器 D 的主要任務是區分幻覺網絡 H 生成的特徵 FH 和深度網絡 Ed 生成的特徵 Fd。

However, as already mentioned, the discriminator is also assigned an auxiliary discriminative task, as detailed in the following.
然而，正如已經提到的，鑑別器也被分配了一個輔助鑑別任務，如下詳述。

The architecture of the networks Ed and H is a mix of 2D and 3D convolutions that process a set of frames, and output a feature vector for every frame t of the input volume, i.e. FtH and Ft d.
網絡 Ed 和 H 的架構是 2D 和 3D 卷積的混合，它們處理一組幀，並為輸入卷的每一幀 t 輸出一個特徵向量，即 FtH 和 Ft d。

This means that each frame have a corresponding feature vector, and these may vary even if sampled from the same video, depending on its dynamics and its position t in the input volume.
這意味著每一幀都有一個相應的特徵向量，即使從同一影像中採樣，這些特徵向量也可能會有所不同，這取決於其動態及其在輸入體積中的位置 t。

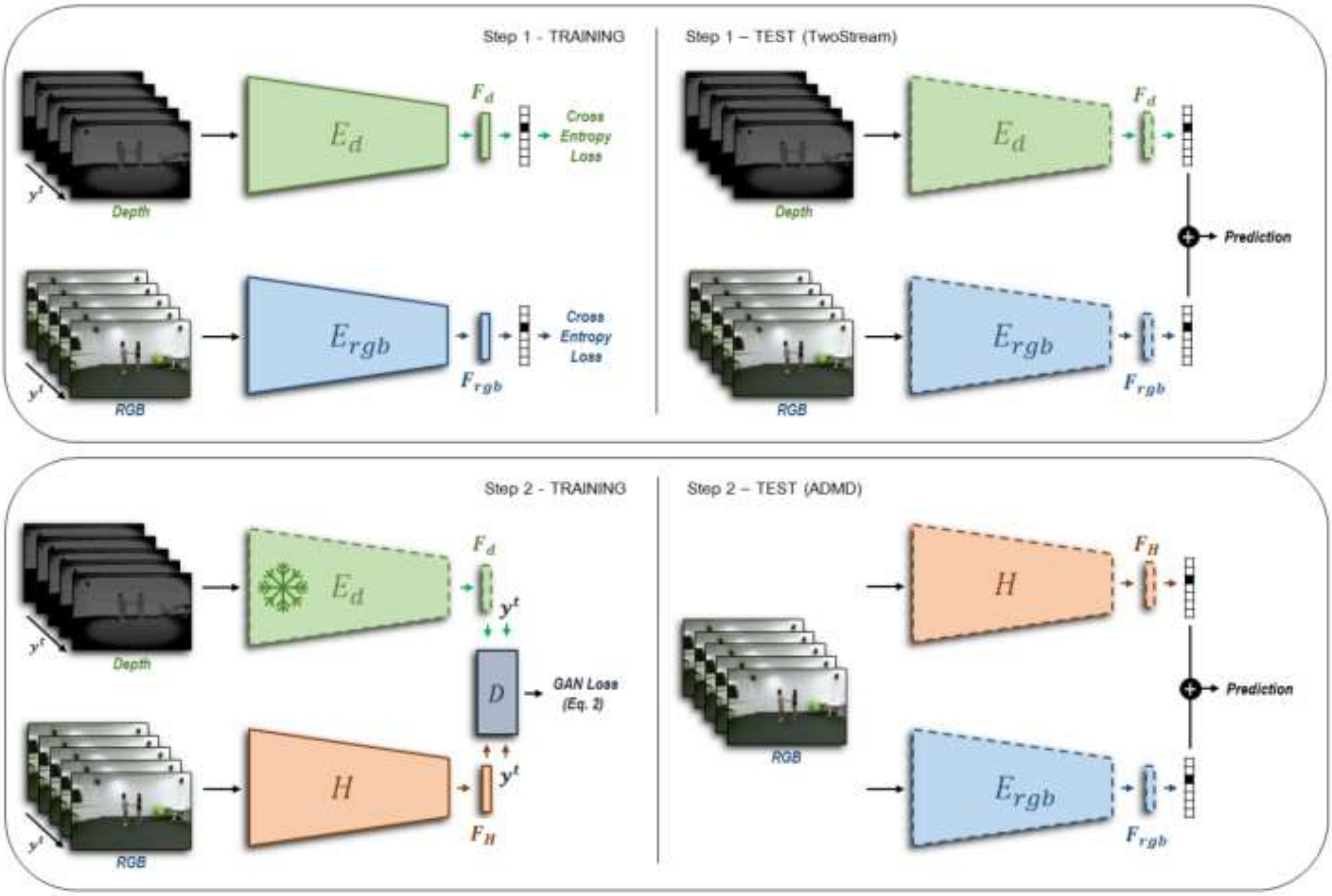
For example, the first frame (and feature vector) of a clip belonging to the action ”shaking hand” might be very different from its the middle frame, but similar to the first frame of a clip belonging to the class ”pushing other person”.
例如，屬於“握手”動作的剪輯的第一幀（和特徵向量）可能與其中間幀非常不同，但類似於屬於“推他人”類的剪輯的第一幀 .

This increases the complexity for the generator, that have not only to generate features similar to Fd, but also to match the order in which they are generated.
這增加了生成器的複雜性，它不僅要生成類似於 Fd 的特徵，還要匹配它們生成的順序。

Namely, FtH should be similar to Ftd, for every frame t of the input volume.
即，對於輸入體積的每一幀 t，FtH 應該類似於 Ftd。

We ease this issue by providing as input to D the one-hot encoding vector of the relative index t, which we

denote yt, concatenated with the respective feature vector, which relates to the CGAN mechanism [46].
我們通過向 D 提供相對索引 t 的 one-hot 編碼向量作為輸入來緩解這個問題，我們表示 yt，與各自的特徵向量連接，這與 CGAN 機制有關 [46]。



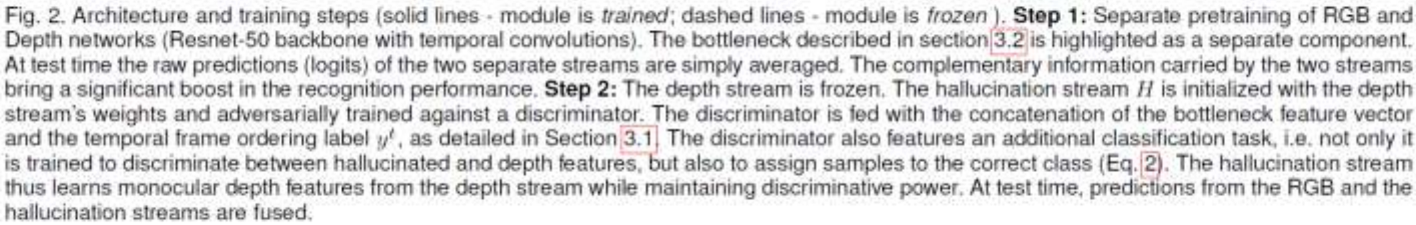


Fig. 2. Architecture and training steps (solid lines - module is *trained*; dashed lines - module is *frozen*). **Step 1:** Separate pretraining of RGB and Depth networks (Resnet-50 backbone with temporal convolutions). The bottleneck described in section 3.2 is highlighted as a separate component. At test time the raw predictions (logits) of the two separate streams are simply averaged. The complementary information carried by the two streams bring a significant boost in the recognition performance. **Step 2:** The depth stream is frozen. The hallucination stream H is initialized with the depth stream's weights and adversarially trained against a discriminator. The discriminator is fed with the concatenation of the bottleneck feature vector and the temporal frame ordering label y^t, as detailed in Section 3.1. The discriminator also features an additional classification task, i.e. not only it is trained to discriminate between hallucinated and depth features, but also to assign samples to the correct class (Eq. 2). The hallucination stream thus learns monocular depth features from the depth stream while maintaining discriminative power. At test time, predictions from the RGB and the hallucination streams are fused.

Fig. 2. Architecture and training steps (solid lines - module is trained; dashed lines - module is frozen).
圖 2. 架構和訓練步驟（實線 - 模塊被訓練；虛線 - 模塊被凍結）。

Step 1: Separate pretraining of RGB and Depth networks (Resnet-50 backbone with temporal convolutions).
第 1 步：單獨預訓練 RGB 和深度網絡（具有時間卷積的 Resnet-50 主幹）。

The bottleneck described in section 3.2 is highlighted as a separate component.
第 3.2 節中描述的瓶頸突出顯示為一個單獨的組件。

At test time the raw predictions (logits) of the two separate streams are simply averaged.

在測試時，兩個獨立流的原始預測（logits）被簡單地平均。

The complementary information carried by the two streams bring a significant boost in the recognition performance.
兩個流攜帶的互補資訊帶來了識別性能的顯著提升。

Step 2: The depth stream is frozen.
步驟 2：深度流被凍結。

The hallucination stream H is initialized with the depth stream’s weights and adversarially trained against a discriminator.
幻覺流 H 用深度流的權重初始化，並針對鑑別器進行對抗訓練。

The discriminator is fed with the concatenation of the bottleneck feature vector and the temporal frame ordering label yt, as detailed in Section 3.1.
鑑別器由瓶頸特徵向量和時間幀排序標籤 yt 的串聯提供，如第 3.1 節所述。

The discriminator also features an additional classification task, i.e. not only it is trained to discriminate between hallucinated and depth features, but also to assign samples to the correct class (Eq. 2).
鑑別器還具有額外的分類任務，即不僅訓練它區分幻覺特徵和深度特徵，而且還將樣本分配給正確的類（等式 2）。

The hallucination stream thus learns monocular depth features from the depth stream while maintaining discriminative power.
因此，幻覺流從深度流中學習單眼深度特徵，同時保持判別能力。

At test time, predictions from the RGB and the hallucination streams are fused.
在測試時，融合了來自 RGB 和幻覺流的預測。

In standard adversarial training, the discriminator D would try to assign the binary label true/fake to the feature vector coming from the two different streams.
在標準對抗訓練中，鑑別器 D 會嘗試將二進制標籤真/假分配給來自兩個不同流的特徵向量。

However, we found that features FH generated with this mechanism, although being very well mixed and indistinguishable from Fd, were struggling to achieve good accuracy for the classification tasks, i.e. were lacking discriminative power.
然而，我們發現用這種機制生成的特徵 FH 儘管混合得很好並且與 Fd 無法區分，但在分類任務中很難達到良好的準確性，即缺乏辨別力。

For this reason we assign to the discriminator the auxiliary task of classifying feature vectors with their correct class.

出於這個原因，我們為鑑別器分配了輔助任務，即用正確的類別對特徵向量進行分類。

The adversarial learning problem is formalized as follows.
對抗性學習問題形式化如下。

Consider the RGB-D dataset (Xrgb, Xd, Y) where xtrgb, xtd (Xrgb, Xd, Y) are time aligned RGB and depth frames, y ~ Y , is the C-dimensional one-hot encoding of the class label, and C is the number of classes for the problem at hand.
考慮 RGB-D 數據集 (Xrgb, Xd, Y)，其中 xtrgb, xtd (Xrgb, Xd, Y) 是時間對齊的 RGB 和深度幀，y ~ Y 是類標籤的 C 維單熱編碼， C 是手頭問題的類數。

Now, let the extended label vector with C + 1 components (classes):
現在，讓擴展標籤向量具有 C+1 個組件（類）：

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

where zeros(C) represents a vector of zeros of dimension C, and || is the concatenation operator.
其中 zeros(C) 表示維度為 C 的零向量，並且 || 是連接運算符。

Using this label vector instead of the classical 0/1 (real/generated) binary label in the discriminator encourages feature representations FH learned by H to encode not only depth (monocular) features, but also to be discriminative.
在鑑別器中使用這個標籤向量而不是經典的 0/1（真實/生成）二元標籤鼓勵 H 學習的特徵表示 FH 不僅編碼深度（單目）特徵，而且具有判別性。

This is possibly why the hallucination network often recovers the accuracy of the teacher and sometimes performs even better, as further discussed in the experimental section.
這可能就是為什麼幻覺網絡經常能恢復老師的準確率，有時甚至表現得更好的原因，如實驗部分進一步討論的那樣。

In summary, we want FH features to be as discriminant as real ones: the adversarial procedure produces fake features which not only are classified as real by the discriminator, but are also assigned to the correct class.
總之，我們希望 FH 特徵與真實特徵一樣具有判別力：對抗性過程產生的假特徵不僅被判別器歸類為真實特徵，而且還被分配到正確的類別。

Based on the above definitions, we define the following minimax game:
基於以上定義，我們定義了以下極大極小博弈：

$$\begin{aligned}\min_{\theta_D}\max_{\theta_H}\ell &= \mathbb{E}_{(x_i,y_i)\sim(X_{rgb},Y)}\mathcal{L}(D(H(x_i)||y^t),\ \hat{y}_i) \\ &+ \mathbb{E}_{(x_i,y_i)\sim(X_d,Y)}\mathcal{L}(|D(E_d(x_i)||y^t),\ \hat{y}_i)\end{aligned} \qquad (2)$$

where θH and θD indicate the parameters of the hallucination stream H and of the discriminator D, || denotes a concatenation operation and L is the softmax cross-entropy function.
其中 θH 和 θD 表示幻覺流 H 和鑑別器 D 的參數，|| 表示連接操作，L 是 softmax 交叉熵函數。

Eq. 2 is optimized via the well known ”label flipping hack” [49], which makes the loss function easier to minimize in practice.
等式 2 通過眾所周知的 ”label flipping hack” [49] 進行優化，這使得損失函數在實踐中更容易最小化。

3.2 Architectural details 建築細節

All three networks (depth stream - Ed, RGB stream - Ergb, and hallucination stream H) are modified Resnet-50 [50] augmented with temporal convolutions and endowed with a final bottleneck layer.
所有三個網絡（深度流 - Ed、RGB 流 - Ergb 和幻覺流 H）都經過修改後的 Resnet-50 [50] 增加了時間卷積並賦予了最終的瓶頸層。

The hallucination networks H are initialized with the respective depth stream weights Ed, following the findings of [10] for object detection, and [11] for action recognition.
幻覺網絡 H 使用各自的深度流權重 Ed 進行初始化，遵循 [10] 用於對象檢測和 [11] 用於動作識別的發現。

Temporal convolutions. 時間卷積

1D temporal convolutions are inserted in the second residual unit of each ResNet layer as illustrated in Fig. 3, following the recent work of Feichtenhofer et al. [19].
在 Feichtenhofer et al. [19] 最近的工作之後，一維時間卷積被插入到每個 ResNet 層的第二個殘差單元中，如圖 3 所示。

For layer l, the weights Wl ∈ R1X1X3XClXCl are convolutional filters initialized as identity mappings at feature level, and centered in time, where Cl is the number of channels in layer l.
對於第 l 層，權重 Wl ∈ R1X1X3XClXCl 是卷積濾波器，在特徵級別初始化為身份映射，並以時間為中心，其中 Cl 是第 l 層中的通道數。

More in detail, all the [1 X 1 X 3] temporal kernels contained in Wl are initialized as [0, 1, 0], i.e. only the information of the central frame is used at the beginning.

更詳細地說，WI 中包含的所有 [1 X 1 X 3] 時間內核都被初始化為 [0, 1, 0]，即在開始時僅使用中心幀的資訊。

This progressively changes as training goes on.
隨著訓練的進行，這種情況會逐漸改變。

Very recently, in [35], the authors explored various network configurations using temporal convolutions, comparing different combinations for the task of video classification.
最近，在 [35] 中，作者使用時間卷積探索了各種網絡配置，比較了影像分類任務的不同組合。

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.
這項工作表明，將 3D 卷積解耦為 2D（空間）和 1D（時間）濾波器是動作識別任務中的最佳設置，可產生最佳精度。

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 [51]).
後一種設置的直覺是在兩個連續的捲積層中分解空間和時間卷積可以簡化空間和時間任務的訓練（也符合 [51]）。

Bottleneck. 瓶頸

Generating, encoding, or aligning high dimensional feature vectors via adversarial training is often a difficult task, due to the inherent instability of the saddle point defined by the GAN minimax game.
由於 GAN 極小極大遊戲定義的鞍點固有的不穩定性，通過對抗訓練生成、編碼或對齊高維特徵向量通常是一項艱鉅的任務。

For this reason, [41] proposes to align a lower dimensional vector, obtained by adding a bottleneck layer to standard architectures.
出於這個原因，[41] 建議對齊一個較低維的向量，這是通過向標準架構添加瓶頸層而獲得的。

This usually does not affect performances of baseline models.
這通常不會影響基線模型的性能。

Indeed, the size of the last ResNet-50 layer (before the logits) is [7, 7, 2048], or simply [2048] after pooling.
實際上，最後一個 ResNet-50 層（在 logits 之前）的大小是 [7, 7, 2048]，或者在池化之後只是 [2048]。

For this reason, we further modify the ResNet-50 by adding either

i) an additional convolutional layer, whose weights Wb ∈ R7X7X2048X128, applied with no padding, reduce

the dimensionality to 128; or ii) a simple 128-dim fully connected layer after pooling.
出於這個原因，我們通過添加任何一個來進一步修改 ResNet-50

i) 一個額外的捲積層，其權重 Wb ∈ R7X7X2048X128，應用無填充，將維度降低到 128； 或者

ii) 池化後的一個簡單的 128 維全連接層。

In Section 4.2.1 we further explore the choice of the bottleneck.
在第 4.2.1 節中，我們進一步探討了瓶頸的選擇。

Input. 輸入

For the task of action recognition, the input to the encoder networks E and H is five 3-channel frames, uniformly sampled from each video sequence, which motivates temporal convolution.
對於動作識別任務，編碼器網絡 E 和 H 的輸入是五個 3 通道幀，從每個影像序列中均勻採樣，這會激發時間卷積。

Instead, for the task of object classification (from single images), no temporal kernels are added to the architecture.
相反，對於對象分類任務（來自單個圖像），架構中沒有添加時間內核。

We try different encodings for the depth channel: for the task of action recognition they are encoded into color images using a jet colormap, as in [52]; for the object recognition task, HHA encoding [53] is already provided in the dataset considered.
我們為深度通道嘗試了不同的編碼：對於動作識別任務，它們使用噴射顏色圖編碼成彩色圖像，如 [52]； 對於物體識別任務，所考慮的資料集中已經提供了 HHA 編碼 [53]。

Discriminator. 鑑別器

The discriminator used to play the adversarial game has different architectures depending on the task.
用於玩對抗遊戲的鑑別器根據任務具有不同的架構。

These architectures follow the empirically validated common practices in the adversarial learning literature, more specifically to what is described in [41].
這些架構遵循對抗性學習文獻中經過經驗驗證的常見實踐，更具體地說是 [41] 中描述的內容。

Its basic structure is that of a multilayer perceptron, stacking fully connected (fc) layers only, since it takes a vector as input (bottleneck features, possibly concatenated with temporal ordering for tasks involving time).
它的基本結構是多層感知器，僅堆疊全連接 (fc) 層，因為它以向量作為輸入（瓶頸特徵，可能與涉及時間的任務的時間順序連接）。

For the task of action recognition, the structure is quite shallow, consisting in D1=[fc(2048), fc(1024), fc(C+1)].
對於動作識別任務，結構相當淺，包括 D1=[fc(2048), fc(1024), fc(C+1)]。

For the task of object classification the structure is instead more complex D2=[fc(1024), fc(1024), fc(1024), fc(2048), fc(3072), fc(C+1)], with skip connections in the lower layers.
對於對象分類的任務，結構反而更複雜 D2=[fc(1024), fc(1024), fc(1024), fc(2048), fc(3072), fc(C+1)], with skip 低層的連接。

Being the former discriminator quite deep, residual connections were inserted in order to allow gradient to flow through the underlying hallucination stream.
作為非常深的前鑑別器，插入了殘差連接以允許梯度流過底層的幻覺流。

Details of the architectures are sketched in Fig. 4.
架構的細節如圖 4 所示。

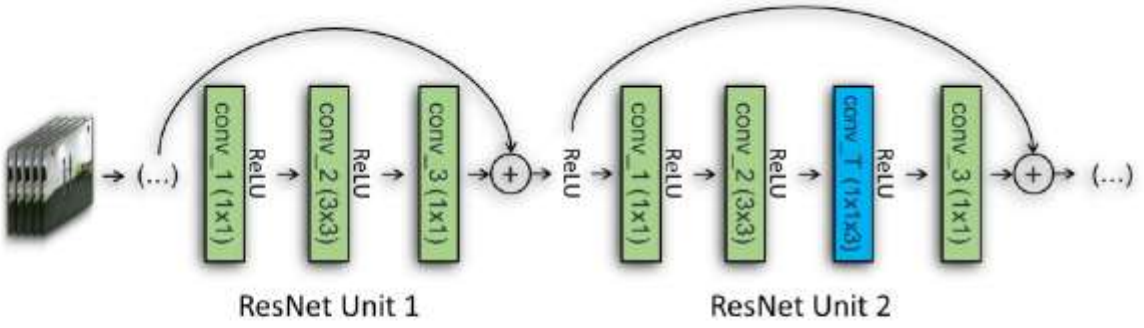


Fig. 3. Detail of the ResNet residual unit with temporal convolutions (blue block).

Fig. 3. Detail of the ResNet residual unit with temporal convolutions (blue block).
圖 3. 帶有時間卷積的 ResNet 殘差單元的細節（藍色塊）。

4 EXPERIMENTS 實驗

4.1 Datasets 資料集

We evaluate the performance of our method on one object classification and two video action classification datasets.
我們評估了我們的方法在一個對象分類和兩個影像動作分類資料集上的性能。

For both tasks the model is initialized with ImageNet pretrained weights.
對於這兩個任務，模型都使用 ImageNet 預訓練權重進行初始化。

For the experiments on the smaller action recognition dataset NW-UCLA, we fine-tune the model starting from the RGB and depth streams trained on the larger NTU RGB+D dataset.
對於在較小的動作識別資料集 NW-UCLA 上的實驗，我們從在較大的 NTU RGB+D 資料集上訓練的 RGB 和深度流開始對模型進行微調。

NTU RGB+D [16].

This is the largest public dataset for multimodal video action recognition.
這是用於多模態影像動作識別的最大公共資料集。

It is composed by 56,880 videos, available in four modalities: RGB videos, depth sequences, infrared frames, and 3D skeleton data of 25 joints (RGB and depth examples illustrated in Fig. 1).
它由 56,880 個影像組成，有四種模式：RGB 影像、深度序列、紅外幀和 25 個關節的 3D 骨架資料（RGB 和深度示例如圖 1 所示）。

It was acquired with a Kinect v2 sensor in 80 different viewpoints, and includes 40 subjects performing 60 distinct actions.
它是使用 Kinect v2 傳感器在 80 個不同視角下獲得的，包括 40 個執行 60 個不同動作的對象。

We follow the two evaluation protocols originally proposed in [16], which are cross-subject and cross-view.
我們遵循 [16] 中最初提出的兩個評估協議，即跨學科和跨視圖。

As in the original paper, we use about 5% of the training data as validation set for both protocols.
與原始論文一樣，我們使用大約 5% 的訓練資料作為兩種協議的驗證集。

The masked depth maps are converted to a three channel map via a jet mapping, as in [52].
蒙版深度圖通過噴射映射轉換為三通道圖，如 [52] 中所示。

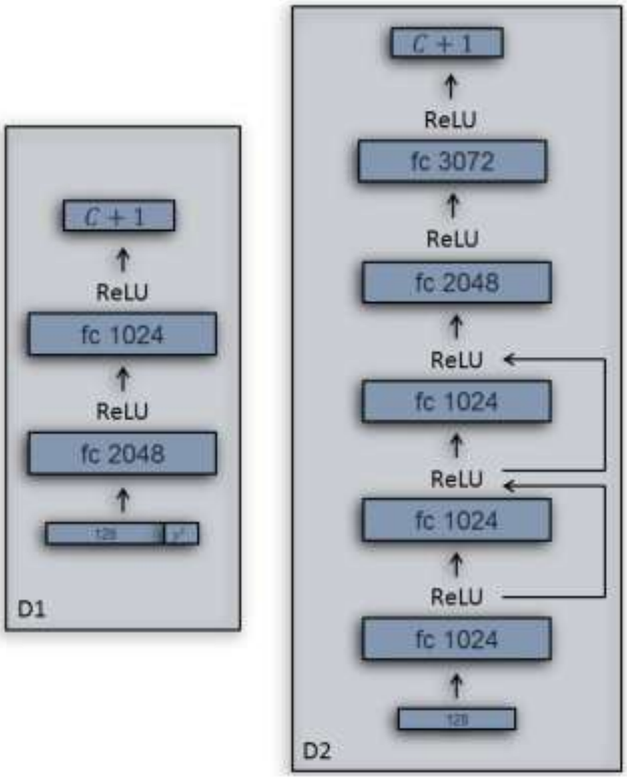


Fig. 4. Architectures for the discriminators used for the two different tasks. Left: D1 for object recognition. Right: D2 for action recognition.

Fig. 4. Architectures for the discriminators used for the two different tasks.
圖 4. 用於兩個不同任務的鑑別器的架構。

Left: D1 for object recognition. Right: D2 for action recognition.
左：用於物體識別的 D1。 右：用於動作識別的 D2。

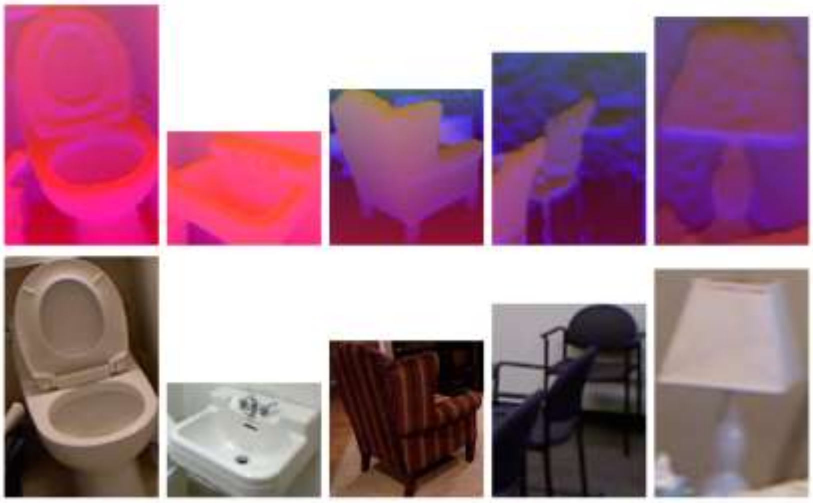


Fig. 5. Examples of RGB and depth frames from the NYUD (RGB-D) dataset.

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

圖 5. 來自 NYUD (RGB-D) 資料集的 RGB 和深度幀示例。

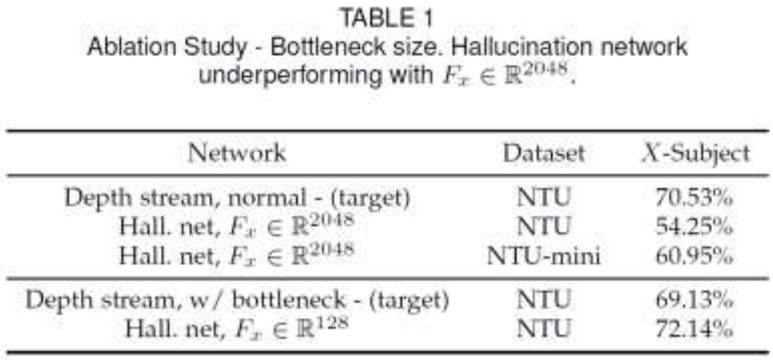


TABLE 1
Ablation Study - Bottleneck size. Hallucination network underperforming with $F_x \in \mathbb{R}^{2048}$.

Network	Dataset	X-Subject
Depth stream, normal - (target)	NTU	70.53%
Hall. net, $F_x \in \mathbb{R}^{2048}$	NTU	54.25%
Hall. net, $F_x \in \mathbb{R}^{2048}$	NTU-mini	60.95%
Depth stream, w/ bottleneck - (target)	NTU	69.13%
Hall. net, $F_x \in \mathbb{R}^{128}$	NTU	72.14%

TABLE 1-表格 1

Ablation Study - Bottleneck size.

消融研究 - 瓶頸尺寸。

Hallucination network underperforming with Fx 2 R2048.

幻覺網絡在 Fx 2 R2048 上表現不佳。

Northwestern-UCLA [17].

This action recognition dataset provides RGB, depth and skeleton sequences for 1475 samples.
這個動作識別資料集提供了 1475 個樣本的 RGB、深度和骨架序列。

It features 10 subjects performing 10 actions captured in 3 different views.
它有 10 個主體執行 10 個動作，在 3 個不同的視圖中捕獲。

NYUD (RGB-D)

This dataset of objects (see examples in Fig. 5) is gathered by cropping out tight bounding boxes around instances of 19 object classes present in the NYUD [15] dataset.
該對像資料集（參見圖 5 中的示例）是通過在 NYUD [15] 資料集中存在的 19 個對像類的實例周圍裁剪出緊密邊界框來收集的。 .

It comprises 2,186 paired labeled training images and 2,401 test images (RGB-D).
它包括 2,186 張成對的標記訓練圖像和 2,401 張測試圖像 (RGB-D)。

Depth images are HHA-encoded [53].
深度圖像是 HHA 編碼的 [53]。

This version of the dataset was proposed in [10] but also used in [41], [54], [55] for the task of modality adaptation, in the framework of domain adaptation (train on one modality, adapt and test the model on the other modality).

這個版本的資料集是在 [10] 中提出的，但也在 [41]、[54]、[55] 中用於模態適應的任務，在域適應的框架內（在一種模態上訓練，適應和測試模型 在另一種方式上）。

The task here is object classification, training on both modalities and testing on RGB only.
這裡的任務是對象分類、兩種模式的訓練和僅在 RGB 上的測試。

4.2 Ablation Study 消融研究

The ablation study is performed on part of the NTU RGB+D dataset, designated as mini-NTU, which consists of random samples from the training set, considering approximately a third of the original dataset size.
消融研究是在 NTU RGB+D 資料集的一部分上進行的，指定為 mini-NTU，它由來自訓練集的隨機樣本組成，考慮到原始資料集大小的大約三分之一。

The test set is still the same as used in the other experiments and defined originally in [16].
測試集仍然與其他實驗中使用的相同，最初在 [16] 中定義。

We study how the hallucination network performance is affected by (1) feeding different types of input to the discriminator, and (2) having the discriminator to perform different tasks.
我們研究了幻覺網絡性能如何受到以下因素的影響：（1）向鑑別器提供不同類型的輸入，以及（2）讓鑑別器執行不同的任務。

4.2.1 Bottleneck size 瓶頸大小

The discriminator receives as input the feature vector FH or Fd from either the hallucination or the depth stream, respectively, along with the frame index label yt.
鑑別器分別接收來自幻覺或深度流的特徵向量 FH 或 Fd 作為輸入，以及幀索引標籤 yt。

It is known that a too big feature vector may cause the GAN training to underperform [41], which we also observe in our experiments, reported in Table 1.
眾所周知，太大的特徵向量可能會導致 GAN 訓練表現不佳 [41]，我們在實驗中也觀察到了這一點，如表 1 所示。

We first trained our depth network without bottleneck on the full NTU dataset, reaching 70.53% accuracy.
我們首先在完整的 NTU 資料集上無瓶頸地訓練了我們的深度網絡，達到了 70.53% 的準確率。

This network is then used as target to learn the hallucination model.
然後將該網絡用作學習幻覺模型的目標。

We observed that the hallucination model trained without bottleneck, i.e., the input to the discriminator is the 2048-dimensional feature vector, is far from recovering the performance of the target (reaching only 54.25%), even if the training space is reduced to the NTU-mini dataset (60.95%).

我們觀察到沒有瓶頸訓練的幻覺模型，即判別器的輸入是 2048 維特徵向量，遠沒有恢復目標的性能（僅達到 54.25%），即使訓練空間減少到 NTU-mini 資料集 (60.95%)。

We then train a network with a 128-dimensional bottleneck (69.13%), initialized with the previous depth stream, except for the bottleneck that is randomly initialized with the MSRA initialization [56].
然後我們訓練一個具有 128 維瓶頸 (69.13%) 的網絡，使用之前的深度流初始化，除了使用 MSRA 初始化隨機初始化的瓶頸 [56]。

The hallucination model that learns using the bottleneck feature vector is able not only to recover, but to surpass the performance of the depth stream, reaching 72.14% accuracy. We observed this behaviour in other experiments along the paper, and we comment that later in Section 4.4.
使用瓶頸特徵向量學習的幻覺模型不僅能夠恢復，而且能夠超越深度流的性能，達到 72.14% 的準確率。 我們在論文中的其他實驗中觀察到了這種行為，我們在後面的 4.4 節中對此進行了評論。

4.2.2 Bottleneck implementation 瓶頸實施

In Table 2 we investigate different ways to decrease the size of Fx from R2048 to R128, as suggested in [41].
在表 2 中，我們研究了將 Fx 的大小從 R2048 減小到 R128 的不同方法，如 [41] 中所建議的。

After the last feature map, which is of dimension 7*7*2048, we tested the three following ways:
在最後一個維度為 7*7*2048 的特徵圖之後，我們測試了以下三種方式：

convolution of [128,7,7] to 1*1*128,
[128,7,7] 卷積到 1*1*128

spatial convolution of [7,7] to 1*1*2048 followed by 1D convolution to 1*1*128, and
，[7,7] 空間卷積到 1*1*2048 然後一維卷積到 1*1*128，和

pooling layer to 1*1*2048 followed by 1D convolution to 1*1*128
池化層到 1*1*2048 然後一維卷積到 1*1*128

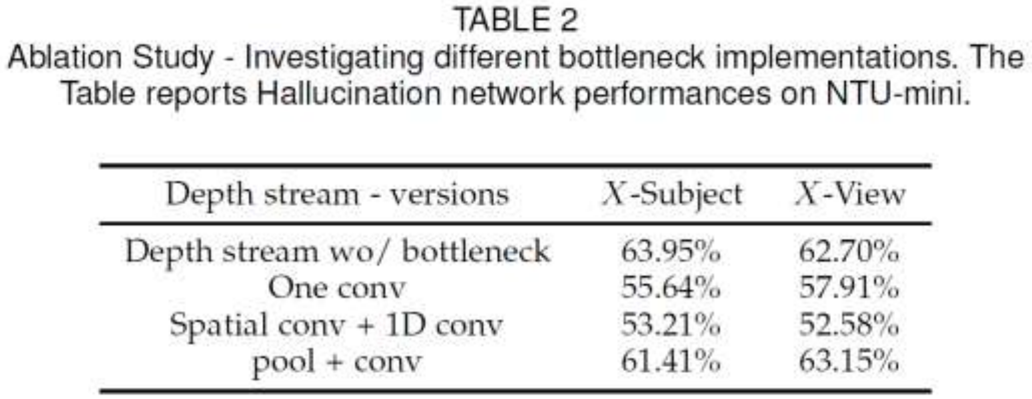


TABLE 2
Ablation Study - Investigating different bottleneck implementations. The Table reports Hallucination network performances on NTU-mini.

Depth stream - versions	X-Subject	X-View
Depth stream wo/ bottleneck	63.95%	62.70%
One conv	55.64%	57.91%
Spatial conv + 1D conv	53.21%	52.58%
pool + conv	61.41%	63.15%

TABLE 2 表 2

Ablation Study - Investigating different bottleneck implementations.

消融研究 - 調查不同的瓶頸實施。

The Table reports Hallucination network performances on NTU-mini.

該表報告了 NTU-mini 上的幻覺網絡性能。

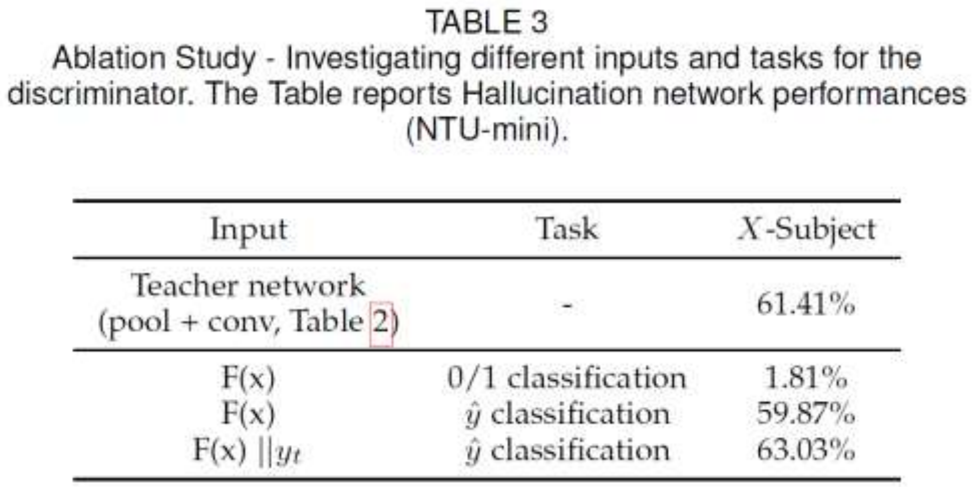


TABLE 3

Ablation Study - Investigating different inputs and tasks for the discriminator. The Table reports Hallucination network performances (NTU-mini).

Input	Task	X-Subject
Teacher network (pool + conv, Table 2)	-	61.41%
F(x)	0/1 classification	1.81%
F(x)	$\hat{y}$ classification	59.87%
F(x) $\|\|y_t$	$\hat{y}$ classification	63.03%

TABLE 3 表 3

Ablation Study - Investigating different inputs and tasks for the discriminator.

消融研究 - 調查鑑別器的不同輸入和任務。

The Table reports Hallucination network performances (NTU-mini).

該表報告了幻覺網絡性能 (NTU-mini)。

Even though the depth stream is just trained on the NTU-mini (63.95% for cross subject, and 62.70% for cross view), the hallucination stream that implements the pool+conv bottleneck is able to recover almost completely (61.41% for cross subject), or even surpass (63.15% for cross view), the original depth stream performance.

即使深度流只是在 NTU-mini 上訓練（跨學科 63.95%，交叉視角 62.70%），實現 pool+conv 瓶頸的幻覺流幾乎可以完全恢復（跨學科 61.41%)，甚至超過（63.15% for cross view），原始深度流性能。

This was the architectural choice we used in the rest of the experiments.

這是我們在其餘實驗中使用的架構選擇。

4.2.3 Discriminator: inputs and tasks 鑑別器：輸入和任務

In this section, we explore whether the task assigned to the discriminator influences the hallucination performance.

在本節中，我們探討分配給鑑別器的任務是否會影響幻覺表現。

As introduced in Section 3, our hypothesis is that the generator has the difficult task of generating features that not only correspond to depth features, but also need to be temporally paired with these.
如第 3 節所述，我們的假設是生成器的艱鉅任務是生成不僅與深度特徵相對應的特徵，而且還需要在時間上與這些特徵配對。

We solve this by introducing the additional information of the frame index yt, which specifies the desired alignment.
我們通過引入幀索引 yt 的附加資訊來解決這個問題，它指定了所需的對齊方式。

Table 3 shows results regarding the
表 3 顯示了關於

(1) traditional binary task of a GAN generator having as input the feature bottleneck,
(1) 以特徵瓶頸作為輸入的 GAN 生成器的傳統二元任務，

(2) the ^y classification task having the same input as before, and
(2) ^y 分類任務具有與之前相同的輸入，以及

(3) the proposed approach.
(3)建議的方法。

The traditional binary task (1) converges to a perfect equilibrium, but the hallucination stream's accuracy is close to random chance, meaning that the learned features are not discriminant at all.
傳統的二元任務 (1) 收斂到一個完美的平衡點，但幻覺流的準確性接近隨機概率，這意味著學習到的特徵根本沒有判別力。

The second approach (2) is able to learn discriminative features, but the addition of the frame order supervision yt (3) shows an increase in performance.
第二種方法 (2) 能夠學習判別特徵，但添加幀順序監督 yt (3) 顯示了性能的提高。

It is reasonable that this mechanism produces maximized gains on more challenging and diverse datasets, as the full NTU dataset, or in fully 3d-convolutional architectures such as I3D [57], due to the higher dependence on temporal convolutions.
由於對時間卷積的更高依賴性，這種機制在更具挑戰性和多樣性的資料集上產生最大化收益是合理的，如完整的 NTU 資料集，或在全 3d 卷積架構（如 I3D [57]）中。

4.3 Action recognition performance and comparisons 動作識別性能及比較

Table 4 compares performances of different methods in the literature, across the two datasets for action recognition - two protocols for the NTU RGB+D and the NW-UCLA.

表 4 比較了文獻中不同方法在動作識別的兩個數據集上的性能—NTU RGB+D 和 NW-UCLA 的兩個協議。

The standard performance measure used for this task and datasets is classification accuracy, estimated according to the protocols, training and testing splits defined in the respective works.
用於此任務和資料集的標準性能度量是分類準確度，根據各自作品中定義的協議、訓練和測試拆分進行估計。

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.
表格的第一部分（用符號表示）指的是無監督方法，即使在學習表示中不依賴標籤，也能取得驚人的高結果。

The second part refers to supervised methods (indicated by 4), divided according to the modalities used for training and testing.
第二部分是監督方法（用 4 表示），根據用於訓練和測試的方式進行劃分。

Here, we report the performance of the separate RGB and depth (with and without bottleneck) streams trained in step 1 (rows #7 and #8).
在這裡，我們報告了在步驟 1（第 7 行和第 8 行）中訓練的單獨 RGB 和深度（有和沒有瓶頸）流的性能。

The small increase in performance is probably due to the extra training steps with small learning rate, after initialized with the bottleneck version trained on the mini-NTU (used for the ablation study).
性能的小幅提升可能是由於在 mini-NTU（用於消融研究）上訓練的瓶頸版本初始化後，額外的訓練步驟和小學習率。

Importantly, the depth stream with bottleneck represents the teacher network used for the hallucination learning.
重要的是，具有瓶頸的深度流代表用於幻覺學習的教師網絡。

We expect our final model to perform better than the one trained on RGB only, whose accuracy constitutes a lower bound for the usefulness of our hallucination model.
我們希望我們的最終模型比僅在 RGB 上訓練的模型表現更好，其準確性構成了我們的幻覺模型有用性的下限。

The values reported for our step 1 models for the NW-UCLA dataset, i.e. the RGB and depth streams, refer to the fine-tuning of our NTU model.
為 NW-UCLA 資料集的第 1 步模型報告的值，即 RGB 和深度流，指的是我們 NTU 模型的微調。

In contrast with [11], and for clearer analysis, the two-stream setup is always not finetuned.

與 [11] 相比，為了更清晰的分析，雙流設置始終沒有進行微調。

Its accuracy represents an upper bound for the final model, which will not rely on depth data at test time.
它的準確度代表了最終模型的上限，它不會依賴於測試時的深度資料。

We have experimented training using pre-trained ImageNet weights instead of the NTU, but it led to lower accuracy.
我們已經嘗試使用預訓練的 ImageNet 權重代替 NTU 進行訓練，但它導致了較低的準確率。

The last part of the table (indicated by □) reports the performance of methods in the privileged information framework, thus directly comparable to ours.
表格的最後一部分（用□表示）報告了特權資訊框架中方法的性能，因此可以直接與我們的進行比較。

The performance values that refer to the Hoffman et al. method [10] (row #20 of Table 4) are taken from the implementation and experiments in [11].
參考 Hoffman 等人的性能值。 方法 [10]（表 4 的第 20 行）取自 [11] 中的實施和實驗。

Row #21 refers to the method by Luo and colleagues [21], that uses 6 modalities at training time (RGB, depth, optical flow, and three different encoding methods for skeleton data), and RGB only at test time.
第 21 行引用了 Luo 及其同事 [21] 的方法，該方法在訓練時使用 6 種模態（RGB、深度、光流和骨架資料的三種不同編碼方法），而在測試時僅使用 RGB。

Step 3 and 4 of [11] (row #22 and #23) refer to the twostream model after the hallucination learning, and its finetuning, respectively.
[11] 的第 3 步和第 4 步（第 22 行和第 23 行）分別參考了幻覺學習及其微調後的雙流模型。

We note that, for simplicity, the results of ADMD Two-Stream models are merely the outcome of the average of the two streams' logits, and are not subject to any fine-tuning, which means that they are directly comparable with row #22.
我們注意到，為了簡單起見，ADMD Two-Stream 模型的結果僅僅是兩個流的 logits 的平均值的結果，並且沒有經過任何微調，這意味著它們可以直接與第 22 行進行比較 .

In addition, results of row #24 correspond to the hallucination stream only.
此外，第 24 行的結果僅對應於幻覺流。

We note that the hallucination stream (row #24) manages to recover and surpass the depth teacher stream (row #8) for the NW-UCLA dataset (83.94% compared to 71.09%), while for the NTU p1 (67.57%) and p2 (71.80%) protocols is around 4% below the respective teacher (71.87% and 75.32%).
我們注意到幻覺流（第 24 行）設法恢復並超過了 NW-UCLA 資料集的深度教師流（第 8 行）（83.94% 與 71.09% 相比），而對於 NTU p1（67.57%）和 p2 (71.80%) 協議比各自的教師（71.87% 和

75.32%）低約 4%。

Nevertheless, when combined with the RGB stream, it performs better (NTU p2 - 81.50%) or comparable (NTU p1 - 73.11%) to the fine-tuned model presented in [11].
然而，當與 RGB 流結合時，它的性能更好 (NTU p2 - 81.50%) 或與 [11] 中提出的微調模型相當 (NTU p1 - 73.11%)。

Since the RGB stream is performing equally well in this work and in [11], we can conclude that the gains in performance are due to better hallucination features.
由於 RGB 流在這項工作和 [11] 中的表現同樣出色，我們可以得出結論，性能的提高是由於更好的幻覺特徵。

4.4 Object recognition performance and comparisons 物體識別性能和比較

Table 5 illustrates the main results obtained for NYUD dataset for the object recognition task.
表 5 說明了 NYUD 資料集在對象識別任務中獲得的主要結果。

As opposed to action recognition, depth information is often noisy here (cfr. Fig. 5 - chair and lamp), probably due to the small resolution of the bounding box crops.
與動作識別相反，這裡的深度資訊通常是嘈雜的（參見圖 5 - 椅子和燈），可能是由於邊界框裁剪的分辨率小。

Depth alone is in fact performing worse than RGB alone (more than 10% gap).
事實上，單獨的深度比單獨的 RGB 表現更差（超過 10% 的差距）。

Still, the amount of complementary information carried by the two modalities is able, when fused in the twostream model, to boost recognition accuracy by more than 5 percentage points, despite the poor depth performance (RGB!52.90%, Depth!40.19%) two-stream!57.39%).
儘管如此，儘管深度性能很差（RGB！52.90%，Depth！40.19%），但當在雙流模型中融合時，兩種模態攜帶的互補資訊量能夠將識別準確度提高 5 個百分點以上 -流！57.39%）。

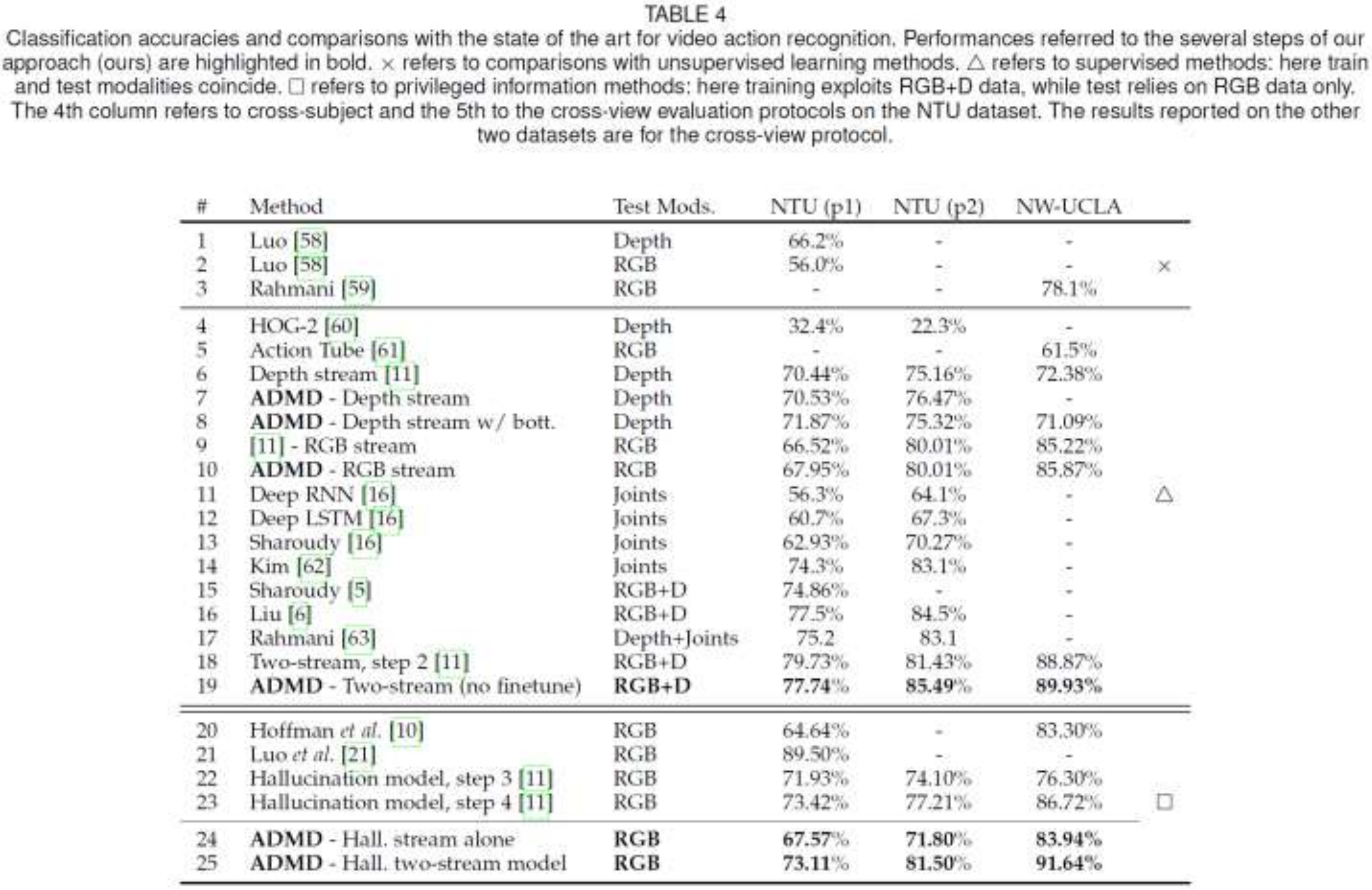


TABLE 4

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%**	

TABLE 4 表 4

Classification accuracies and comparisons with the state of the art for video action recognition.
影像動作識別的分類精度和與現有技術的比較。

Performances referred to the several steps of our approach (ours) are highlighted in bold.
涉及我們方法（我們的）的幾個步驟的性能以粗體突出顯示。

refers to comparisons with unsupervised learning methods.
指與無監督學習方法的比較。

4 refers to supervised methods: here train and test modalities coincide.
4 指的是監督方法：這裡訓練和測試模式一致。

refers to privileged information methods: here training exploits RGB+D data, while test relies on RGB data only.
指特權資訊方法：這裡的訓練利用 RGB+D 資料，而測試僅依賴於 RGB 資料。

The 4th column refers to cross-subject and the 5th to the cross-view evaluation protocols on the NTU dataset.
第 4 列涉及跨學科，第 5 列涉及 NTU 資料集上的跨視圖評估協議。

The results reported on the other two datasets are for the cross-view protocol.
在其他兩個資料集上報告的結果是針對跨視圖協議的。

TABLE 5 表 5
Object Recognition 物體識別

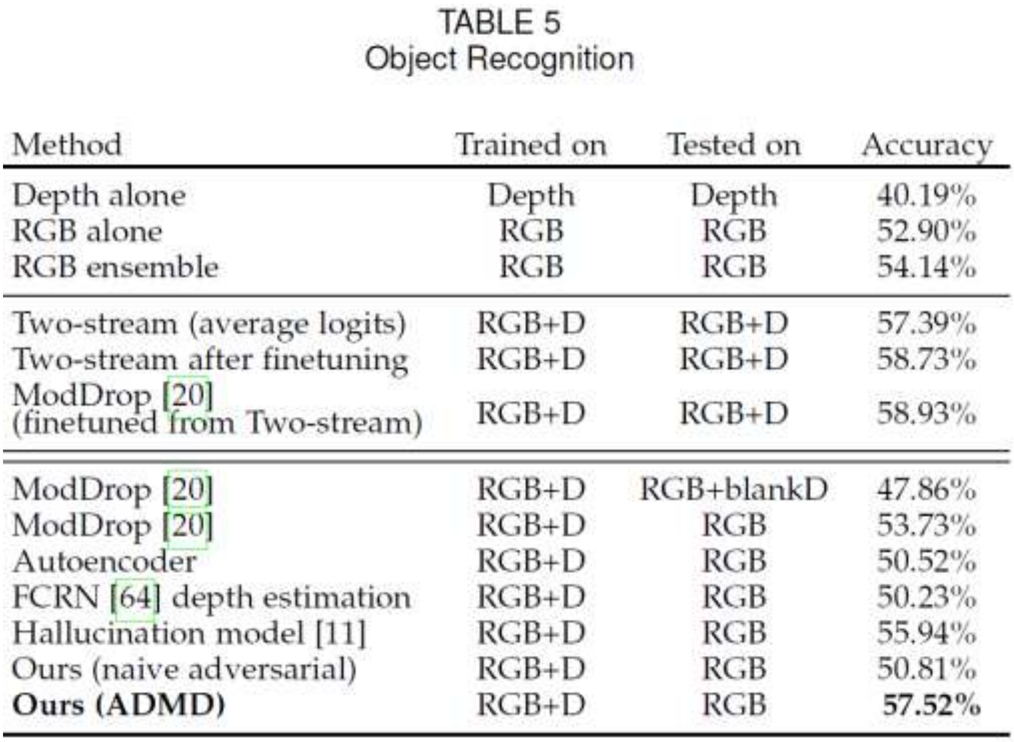


TABLE 5
Object Recognition

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 logits)	RGB+D	RGB+D	57.39%
Two-stream after finetuning	RGB+D	RGB+D	58.73%
ModDrop [20] (finetuned from Two-stream)	RGB+D	RGB+D	58.93%
ModDrop [20]	RGB+D	RGB+blankD	47.86%
ModDrop [20]	RGB+D	RGB	53.73%
Autoencoder	RGB+D	RGB	50.52%
FCRN [64] depth estimation	RGB+D	RGB	50.23%
Hallucination model [11]	RGB+D	RGB	55.94%
Ours (naive adversarial)	RGB+D	RGB	50.81%
Ours (ADMD)	RGB+D	RGB	**57.52%**

It is well established that ensemble methods tend to outperform their single-model counterparts: an ensemble of two CNNs, each trained started from a different initialization, outperforms either independent model [65].
眾所周知，集成方法往往優於其單一模型對應物：兩個 CNN 的集成，每個 CNN 從不同的初始化開始訓練，優於任一獨立模型 [65]。

Since, in principle, the proposed ADMD strategy is the combination of an RGB model trained using a standard supervised approach and another adversarially trained RGB model, we additionally compare our approach to an ensemble of RGB classifiers (third line of Table 5).
由於原則上提出的 ADMD 策略是使用標準監督方法訓練的 RGB 模型和另一個對抗訓練的 RGB 模型的組合，因此我們另外將我們的方法與 RGB 分類器的集合進行比較（表 5 的第三行）。

Interestingly, despite starting from a two relatively high single-stream performances, the fusion process of two RGB networks only marginally increases the final accuracy (RGB1!53.19%, RGB2!52.60%) Ensemble!54.14%).
有趣的是，儘管從兩個相對較高的單流性能開始，兩個 RGB 網絡的融合過程只是略微提高了最終的準確率（RGB1！53.19%，RGB2！52.60%）Ensemble！54.14%）。

As noticed for the task of action recognition, we found that fine-tuning the fused streams does not always bring significant improvements, as opposed to [11], were the architecture features cross-stream multiplier connections, which need to be trained in an further step.
正如在動作識別任務中註意到的那樣，我們發現微調融合流並不總是帶來顯著的改進，與 [11] 相反，架構具有跨流乘法器連接，需要進一步訓練 步。

Fine-tuning with the strategy proposed by Neverova et al. [20] looks slightly more effective, since ModDrop introduces a light dropout at the input layers, both on the images and on the whole modalities.
使用 Neverova et al. [20]提出的策略進行微調看起來稍微更有效，因為 ModDrop 在輸入層引入了光衰減，包括圖像和整個模態。

The resulting model is tested in both the original setup proposed in [20], namely by blanking out the depth stream, and by simply using RGB predictions.
結果模型在 [20] 中提出的原始設置中進行了測試，即通過消隱深度流和簡單地使用 RGB 預測。

The latter scheme slightly improves the performance of the RGB stream, possibly thanks to dropout.
後一種方案略微提高了 RGB 流的性能，這可能要歸功於 dropout。

However, although the model shows more robustness to missing depth at test time, it clearly fails to extract any monocular depth cue.
然而，儘管該模型在測試時對缺失的深度表現出更強的魯棒性(robustness)，但它顯然無法提取任何單眼深度線索。

Another interesting comparison we perform is the following: we train a cross-modal autoencoder with an L2 loss in order to reconstruct depth maps from RGB.
我們執行的另一個有趣的比較如下：我們訓練一個具有 L2 損失的跨模態自動編碼器，以便從 RGB 重建深度圖。

The encoderdecoder architecture consists in the very same RGB ResNet-50 for the encoder, and in 5 stacked deconvolutional blocks intertwined with batch-norm layers for the decoder.
編碼器解碼器架構包括用於編碼器的相同 RGB ResNet-50，以及用於解碼器的與批規範層交織在一起的 5 個堆疊解卷積塊。

At test time, when depth is not available, we provide RGB frames to the autoencoder, which reconstructs the missing modality to feed the corresponding branch of the two-stream architecture.
在測試時，當深度不可用時，我們向自動編碼器提供 RGB 幀，自動編碼器重建缺失的模態以饋送雙流架構的相應分支。

The performance of this setup is quite poor.
這種設置的性能很差。

We observe that the autoencoder easily overfits the training set, generating high quality depth maps for the training set, while it performs very poorly for the test set.
我們觀察到自動編碼器很容易過擬合訓練集，為訓練集生成高質量的深度圖，而它在測試集上的表現卻很差。

Similarly, we reconstruct depth by means of FCRN [64], a state-of-the-art depth estimator trained on the entire NYUD dataset.
同樣，我們通過 FCRN [64] 重建深度，這是一種在整個 NYUD 資料集上訓練的最先進的深度估計器。

Again, performance is quite poor, since depth estimated by FCRN misses many fine details needed for object classification.
同樣，性能很差，因為 FCRN 估計的深度錯過了對象分類所需的許多精細細節。

This suggest that, for the recognition task, hallucinating task specific features is more effective than estimating depth.
這表明，對於識別任務，幻覺任務特定特徵比估計深度更有效。

This claim is again confirmed with the result for the Hallucination model proposed in [11], adapted in this case for object recognition (Table 5, 3rd to last row).
[11] 中提出的幻覺模型的結果再次證實了這一說法，在這種情況下適用於對象識別（表 5，第 3 行到最後一行）。

This method outperforms both the RGB stream and the RGB ensemble, confirming the value of hallucinating depth. It also outperforms the other baselines that use RGB only at test time (3rd section of Table 5).
這種方法優於 RGB 流和 RGB 集成，證實了幻覺深度的價值。 它還優於僅在測試時使用 RGB 的其他基線（表 5 的第 3 部分）。

In particular, it performs considerably better than FCRN depth estimation, which indicates again that depth feature hallucination is more effective than predicting depth maps at pixel level.
特別是，它的性能比 FCRN 深度估計要好得多，這再次表明深度特徵幻覺比在像素級別預測深度圖更有效。

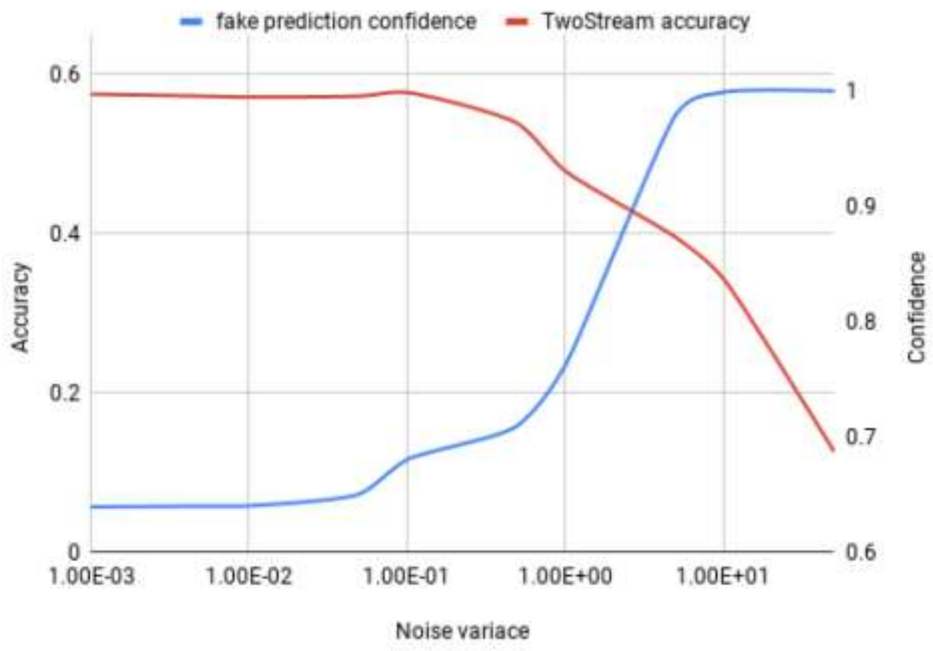
More importantly, we can directly compare it with ADMD proposed in this paper (55.94% vs 57.52%), concluding that, similarly to action recognition experiments, the adversarial approach performs better.
更重要的是，我們可以直接將其與本文提出的 ADMD 進行比較（55.94% vs 57.52%），得出的結論是，與動作識別實驗類似，對抗性方法表現更好。

Eventually, we tested our adversarial scheme in two different setups: i) the naive setup where the discriminator D is assigned the binary task only, and ii) the ADMD setup, where the discriminator is also assigned the classification task.
最終，我們在兩種不同的設置中測試了我們的對抗方案：i) 天真設置，其中鑑別器 D 僅被分配了二元

任務，以及 ii) ADMD 設置，其中鑑別器也被分配了分類任務。

While the former performs as the autoencoder, the latter is able to fully recover the accuracy of the Two-stream model, being only slightly below that of the fine-tuned model.
前者作為自動編碼器，後者能夠完全恢復雙流模型的精度，僅略低於微調模型的精度。

4.5 Inference with noisy depth 帶有噪聲深度的推理



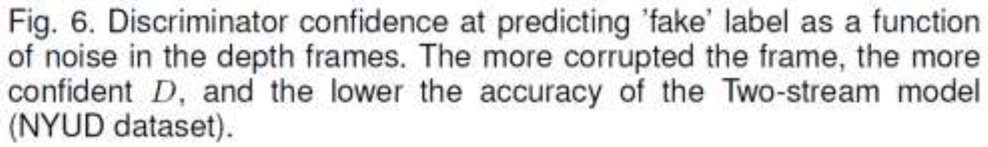


Fig. 6. Discriminator confidence at predicting 'fake' label as a function of noise in the depth frames. The more corrupted the frame, the more confident D, and the lower the accuracy of the Two-stream model (NYUD dataset).

Fig. 6. Discriminator confidence at predicting ’fake’label as a function of noise in the depth frames.
圖 6. 預測“假”標籤作為深度幀中噪聲函數的鑑別器置信度

The more corrupted the frame, the more confident D, and the lower the accuracy of the Two-stream model (NYUD dataset).
幀損壞越多，D 越有信心，雙流模型（NYUD 資料集）的準確度就越低。

In real test scenarios, it is often the case that we can only access noisy depth data.
在真實的測試場景中，我們常常只能訪問嘈雜的深度資料。

In this section, we address two questions:
在本節中，我們將解決兩個問題：

i) how much such noisy data can degrade the performance of a multimodal setup?
i) 這種嘈雜的資料會在多大程度上降低多模式設置的性能？

ii) At which level of noise does it become favorable to hallucinate the depth modality with respect to using the teacher model (Twostream) with noisy depth data?
ii) 相對於使用帶有噪聲深度資料的教師模型 (Twostream) 而言，在哪個噪聲水平上對深度模態產生幻覺更有利？

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 [66].
NTU 資料集 (Kinect) 中使用的深度傳感器是一個 IR 發射器，與一個 IR 攝像頭耦合，並且具有非常複雜的噪聲表徵，包括至少 6 個不同的源 [66]。

It is beyond the scope of this work to investigate noise models affecting the depth channel, so, for our analysis we choose the most influencing one, i.e., multiplicative speckle noise.
研究影響深度通道的噪聲模型超出了這項工作的範圍，因此，對於我們的分析，我們選擇影響最大的模型，即乘法散斑噪聲。

Hence, we inject Gaussian noise in the depth images I in order to simulate speckle noise:
因此，我們在深度圖像 I 中註入高斯噪聲以模擬散斑噪聲：
I = I * n, n ~ N(1, σ)

Table 6 shows how performances of our Two-stream network degrade when depth is corrupted with such Gaussian noise with increasing variance (NTU cross-view protocol and NYUD).
表 6 顯示了當深度被這種隨著方差增加的高斯噪聲破壞時，我們的雙流網絡的性能如何下降（NTU 交叉視圖協議和 NYUD）。

Results show that accuracy significantly decreases with respect to the one guaranteed by our hallucination model (81.50% - row #25) in Table 4, even with low noise variance of σ^2=10^-1.
結果表明，相對於表 4 中我們的幻覺模型（81.50% - 第 25 行）保證的精度，即使 σ^2=10^-1 的低噪聲方差，精度也顯著降低。

For the task of object recognition, we can see that ModDrop [20] is slightly more resilient to depth corruption than the simple Two-stream, since fine-tuned with noise (dropout) in the input layer.
對於對象識別的任務，我們可以看到 ModDrop [20] 比簡單的 Two-stream 更能適應深度損壞，因為在輸入層中用噪聲（dropout）進行了微調。

This experiment shows, in conclusion, that ADMD is able not only deal with a missing modality, but also with a noisy one.
總之，該實驗表明 ADMD 不僅能夠處理缺失的模態，還能夠處理嘈雜的模態。

In an online scenario, the discriminator D, trained in step 2, can give an indication on when to operatively switch from Two-stream to ADMD, that is, when to substitute the depth branch with the hallucination.
在在線場景中，在步驟 2 中訓練的鑑別器 D 可以指示何時可操作地從雙流切換到 ADMD，即何時用幻覺代替深度分支。

When training reaches equilibrium, D is maximally fooled by the features generated by H, and cannot distinguish them from those encoded by Ed.
當訓練達到平衡時，D 被 H 生成的特徵最大限度地愚弄，無法將它們與 Ed 編碼的特徵區分開來。

In practice, this means that the predicted probability for the fake class (last class in ^y, eq. 1) is p(^y = C + 1) ≈ .5 on average.
在實踐中，這意味著假類（^y 中的最後一個類，方程 1）的預測概率平均為 p(^y = C + 1) ≈ .5。

However, when features computed from corrupted depth start to flow inside D, its prediction for the fake class starts to be more and more confident.
然而，當從損壞的深度計算的特徵開始流入 D 內部時，它對假類的預測開始越來越有信心。

Figure 6 plots the behavior of D as noise increases, together with accuracy of the Two-stream model.
圖 6 繪製了 D 隨著噪聲增加的行為以及雙流模型的準確性。

There is a clear turning point in both accuracy and confidence, which can be employed in practice to decide when to switch from Ed to H i.e. when to drop depth as a modality and start using monocular depth features extracted from RGB.
準確度和置信度都有一個明顯的轉折點，可以在實踐中決定何時從 Ed 切換到 H，即何時將深度作為一種模式並開始使用從 RGB 中提取的單眼深度特徵。

4.6 Discussion 討論

Some interesting points arise from the analysis of our findings, which we summarize in the following.
一些有趣的觀點來自對我們發現的分析，我們在下面總結。

1. RGB and depth actually carry complementary information.
RGB 和深度實際上攜帶補充資訊。

As a matter of fact, the Two-stream setup always provides a surprisingly better accuracy than the two streams alone.
事實上，雙流設置總是比單獨使用兩個流提供令人驚訝的更好的準確性。

As additional evidence, a multimodal ensemble (i.e. the Two-stream) performs better than a mono-modal ensemble (Table 5), despite the lower accuracy of one of its singlestream components (either depth or RGB,

depending on task and dataset).
作為額外的證據，儘管其單流組件之一（深度或 RGB，取決於任務和資料集）的準確性較低，但多模態集合（即雙流）的性能優於單模態集合（表 5）。

2. There is (monocular) depth information in RGB images.
2. RGB 圖像中有（單目）深度資訊。

This is evident from the fact that the hallucination stream often recovers and sometimes surpasses the accuracy of its depth-based teacher network.
從幻覺流經常恢復並且有時超過其基於深度的教師網絡的準確性這一事實中可以明顯看出這一點。

Besides, fusing hallucination and RGB streams always bring the benefits, as fusing RGB and Depth.
此外，融合幻覺和 RGB 流總是帶來好處，如融合 RGB 和深度。

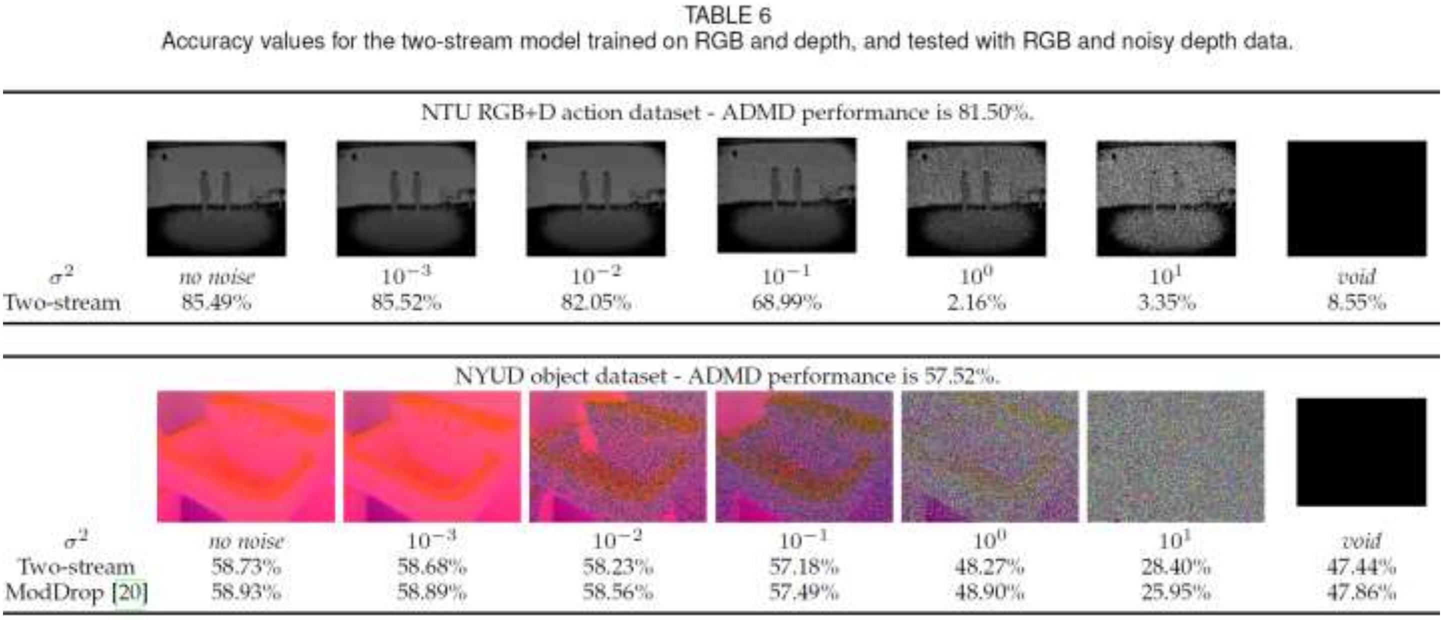


TABLE 6
Accuracy values for the two-stream model trained on RGB and depth, and tested with RGB and noisy depth data.

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%

TABLE 6
表 6

Accuracy values for the two-stream model trained on RGB and depth, and tested with RGB and noisy depth data.
在 RGB 和深度上訓練的雙流模型的準確度值，並使用 RGB 和噪聲深度資料進行測試。

3. Standard supervised learning has limitations in extracting information.
3. 標準監督學習在提取資訊方面存在局限性。

In fact, given the evidence that there is depth information to exploit in RGB images, minimizing cross-entropy loss is not enough to fully extract it.

事實上，鑑於有證據表明 RGB 圖像中存在可利用的深度資訊，最小化交叉熵損失並不足以完全提取它。

For that we need a student-teacher adversarial framework.
為此，我們需要一個學生-教師對抗框架。

This has an interesting parallel in adversarial network compression [38], where the performance of a fully supervised small network can be boosted by adversarial training against a high-capacity (and better performing) teacher net.
這在對抗性網絡壓縮 [38] 中有一個有趣的相似之處，其中完全監督的小型網絡的性能可以通過對抗高容量（和更好性能）教師網絡的對抗性訓練來提高。

In [38], it is also observed that the student can surpass the teacher in some occasions.
在[38]中，還觀察到學生在某些情況下可以超越老師。

4. Adversarial training alone only is not enough.
4. 僅靠對抗訓練是不夠的。

The naive discriminator trained for the binary task (real/generated) is not sufficient to force the hallucination network to produce discriminative features.
為二元任務（真實/生成）訓練的樸素判別器不足以迫使幻覺網絡產生判別特徵。

The auxiliary discriminative task is necessary to extract monocular depth cues which are also discriminative for a given task (on the other hand, the auxiliary task only is not enough, as suggested by the performance of the RGB ensemble).
輔助判別任務對於提取對給定任務也具有判別力的單眼深度線索是必要的（另一方面，僅輔助任務是不夠的，正如 RGB 集成的性能所暗示的那樣）。

5. Hallucinating task-specific depth features is more effective than estimating full depth images.
5. 幻覺特定於任務的深度特徵比估計全深度圖像更有效。

Not only estimated depth is often missing details needed for classification, but also its estimation is driven by mere reconstruction objectives.
不僅估計的深度通常缺少分類所需的細節，而且其估計僅由重建目標驅動。

On the contrary, feature hallucination addresses a specific classification task and requires estimating low dimensional vectors instead of images.
相反，特徵幻覺解決了特定的分類任務，需要估計低維向量而不是圖像。

5 CONCLUSIONS 結論

In this work, we have introduced a novel technique to exploit additional information, in the form of depth images at training time, to improve RGB only models at test time.
在這項工作中，我們引入了一種新技術來利用額外資訊，在訓練時以深度圖像的形式，在測試時改進僅 RGB 模型。

This is done by adversarially training a hallucination network which learns from a teacher depth stream how to encode monocular depth features from RGB frames.
這是通過對抗性訓練幻覺網絡來完成的，該網絡從教師深度流中學習如何從 RGB 幀編碼單眼深度特徵。

The proposed approach outperforms previous ones in the privileged information scenario in the tasks of object classification and action recognition on three different datasets.
在三個不同資料集的對象分類和動作識別任務中，所提出的方法在特權資訊場景中優於以前的方法。

Additionally, the hallucination framework is shown to be very effective in cases where depth is noisy.
此外，幻覺框架在深度嘈雜的情況下非常有效。

Code is available at
https://github.com/pmorerio/admd