

會跳格的小窗口

本課講 2015 年的一篇論文，2016 年在一個大會上正式發表。科學家 Fisher Yu 和 Vladlen Koltun 讓電腦的小窗口「跳著看」，給圖片上每一個點都標對名字。

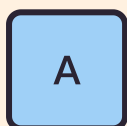
適合 10 歲左右

約 40 分鐘

需要：鉛筆、方格紙、卡紙、剪刀、一張舊雜誌照片、一位同伴

1 熱身：這一小块是什麼？

下面 4 個小方塊，是從 4 張照片裡剪下來的。它們可能是什麼？每塊寫出 2 種猜測。



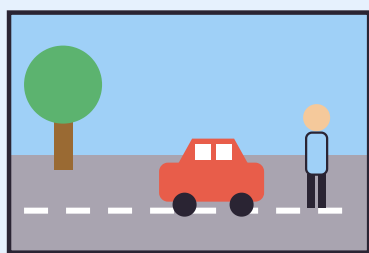
A : _____ B : _____

C : _____ D : _____

藍色的 A，可能是天空，也可能是大海、藍牆、藍衣服。只看一小塊，**誰也說不準**。要是能看到它周圍的東西，就容易多了。

2 比「認出來」更難的事：給每個點塗顏色

第 8 課裡，電腦學會了看一張圖，說出「這是一隻貓」。整張圖只要一個答案。還有一件更難的事：給圖上**每一個小點**都標出它屬於什麼，是天空、馬路、人還是車。這件事叫**語義分割**。「語義」就是「意思」，「分割」就是「分成一塊一塊」。



原來的照片



電腦要交的「塗色圖」

■ 天空 ■ 馬路 ■ 樹 ■ 車 ■ 人

示意圖：右邊每種東西只用一種顏色，車輪也算車。這很像填塗色書。可是照片上沒有「1 號塗藍色」這樣的號碼，電腦得自己判斷。

3 只看一小塊，分不清

仔細看上一節的照片。那個人穿著一件**淺藍色的衣服**，和天空的顏色一模一樣。

電腦看到的範圍	它會怎麼想
只看衣服上的一個點	「一片淺藍……是天空吧？」（猜錯了）
看到點的周圍一小圈	「還是淺藍，有一條黑邊。說不準。」
看到周圍一大片	「上面有頭，下面有腿，這是人的衣服！」（猜對了）

所以，要標對一個點，電腦必須看到它**周圍很大一片**。一個點能看到的周圍範圍，我們叫它**視野**，科學家叫它「感受野」。視野越大，越不容易認錯。

4 老辦法：縮小再放大

第 8 課講過，電腦看圖用一個**小窗口**掃描。窗口一次只看一小塊，比如 3×3 共 9 個格子。窗口小，視野就小。以前的程序有一個辦法：**把圖片一次次縮小**。圖變小了，同樣大的窗口就蓋住了更多內容，就像站遠一點看畫。

原圖：細杆子、車輪

縮小

再放大：細節沒了

可是縮小會丟東西。最後要給每個點塗顏色，圖又得放大回來。丟掉的細節，放大也找不回來。

示意圖：格子越黑，東西越多。縮小時，4 個格子併成 1 個。細杆子變成又寬又淡的帶子，車輪也看不出來了。

想看得遠，就要縮小。想保住細節，就不能縮小。兩位作者想兩樣都要。

5 新辦法：讓窗口跳著看

他們的辦法是：窗口還是 9 個格子，但格子之間**留出空隙**，跳著看。「間隔」表示從一個格子到下一個格子要走幾步。

間隔 1：格子挨著
看到 3×3 一片

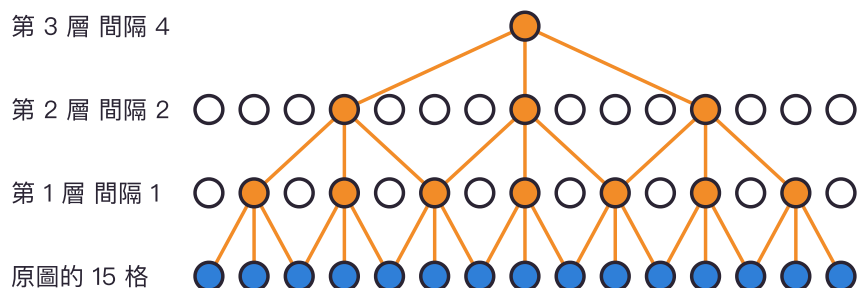
間隔 2：中間空 1 格
看到 5×5 一片

間隔 4：中間空 3 格
看到 9×9 一片

橙色是窗口真正看的 9 個格子，藍框是它蓋住的範圍。

三個窗口都只看 9 個格子，花的力氣一樣多。間隔越大，蓋住的範圍越大。這種跳著看的窗口，叫**空洞卷積**。「卷積」就是第 8 課那種小窗口掃描，「空洞」就是窗口裡留了洞。

一層比一層跳得遠。只跳一次，中間的格子會漏掉。所以他們把窗口疊成好幾層，間隔一層翻一倍：1、2、4、8……下面用一排格子來看（示意圖）：



疊幾層	挨著看	跳著看
1 層	3 格	3 格
2 層	5 格	7 格
3 層	7 格	15 格
4 層	9 格	31 格

最上面那 1 個點，通過 3 層，連到了原圖一整排 15 個格子，一個也沒漏。右邊的表裡，挨著看每多一層只多看 2 格，跳著看每多一層視野差不多翻一倍。而且圖片一直沒有縮小，細節都還在。

6 上下文模塊：專門看大範圍的小組

「上下文」就是一個東西周圍的情況。作者用跳著看的窗口，搭了一個專門看大範圍的小組，叫**上下文模塊**。它一共 8 層，中間幾層的間隔是 1、2、4、8、16。程序先用「認圖部分」給每個點一個初步答案，上下文模塊再看看周圍一大片，改掉錯的。比如，認圖部分把衣服上的點標成了「天空」。上下文模塊一看，這個點上面是頭，下面是腿，就改成「人」（我們編的例子）。

- ・ 剛裝上時什麼都不改。上下文模塊一開始原樣交出答案，再慢慢學會改錯，就不會把對的答案改壞。
- ・ 拿掉兩次縮小。作者從一個認圖程序裡拿掉兩次「縮小」，換成跳著看的窗口。只做這一步，就比當時兩種有名的方法標得更準。加上上下文模塊，又準了一些。在城市街道的照片上，它也好用。

7 這和今天有什麼關係

「跳著看」很快跑到了別的地方。聲音是一長串很密的數字。2016 年，WaveNet 程序用「跳著聽」生成說話的聲音，間隔也是一層翻一倍。2020 年以後，有些讀長文章的 AI 也借用了「跳著看」。

說實話，今天給圖片塗顏色的 AI，很多已經改用「注意力」（第 13 課會講），每個點能一下子看到整張圖。不過，跳著看的窗口還在一些程序裡用著。它的道理也沒過時：**不一定要每個格子都看，聰明地跳著看，也能看得又遠又清楚。**

8 動手：做三張跳格卡片

玩法（1—2 人，約 15 分鐘）

1. 在 3 張卡紙上各畫 9 個小方孔，排成 3 行 3 列，每個孔和方格紙的 1 格一樣大。A 的孔挨著（間隔 1），B 的孔之間空 1 格（間隔 2），C 的孔之間空 3 格（間隔 4）。請大人幫忙剪孔。
2. 把卡片放在方格紙上，透過每個孔點一個點。拿開卡片，數一數點了幾個點，最外面的點圍成的正方形一邊有幾格。
3. 先猜間隔 8 的卡片 D 會蓋住多大，再做一張驗證。

卡片	間隔	點了幾個點	蓋住的範圍（幾格 × 幾格）
A	1		
B	2		
C	4		

你會發現：每張卡片都只點 9 個點，間隔越大，蓋住的範圍越大。點和點中間漏掉的格子，就是電腦要一層層疊起來的原因。

9 動手：只露一小塊，猜猜是什麼

玩法（2—3 人，約 10 分鐘）

1. 準備 2 張舊雜誌照片和 3 張白紙。白紙中間各剪一個方孔：小孔 1 釐米，中孔 3 釐米，大孔 6 釐米。
2. 出題人選照片上的一個點，用小孔的紙蓋住照片，孔對準這個點。猜的人說：這個點是什麼東西的一部分？
3. 換中孔、大孔再猜，孔對準同一個點。最後拿開紙核對。換照片，換人出題。

照片	小孔猜	中孔猜	大孔猜	正確答案
1				
2				

你會發現：看到的周圍越多，越容易猜對。電腦給每個點塗顏色，也需要很大的視野。

10 詞語小卡片

語義分割

給圖片上每一個點都標出它屬於什麼，就像給塗色書的每一塊塗對顏色。

視野（感受野）

一個點能看到的周圍範圍。視野越大，越不容易認錯。

空洞卷積

格子之間留空、跳著看的小窗口。花一樣的力氣，看到更大一片。

上下文模塊

用跳著看的窗口搭成的小組，專門看大範圍，幫忙改掉標錯的點。

11 想一想

1. 熱身時，哪個小方塊最難猜？要看到周圍的什麼東西，你才能猜對？

2. 一輛會自己開的車在看路。為什麼只說「前面有人」不夠，還要知道每個點是馬路還是人？

3. 跳著看會漏掉中間的格子。為什麼一層一層疊起來，就不怕漏了？試著用第 5 節的圖說一說。

一句話記住：窗口跳著看，一層比一層跳得遠，圖片不用縮小，也能看到一大片。

給家長和老師：本課講的是 Fisher Yu 和 Vladlen Koltun 的論文 [Multi-Scale Context Aggregation by Dilated Convolutions](#) (arXiv 2015, ICLR 2016)。論文用空洞卷積 (dilated convolution) 讓感受野隨層數指數增長，同時不降分辨率、不增參數。它設計了可接在分割網絡後面的上下文模塊（空洞率 1、1、2、4、8、16、1，再接 1×1 卷積，恆等初始化），並去掉 VGG-16 最後兩個池化層作為前端。前端在 Pascal VOC 2012 上已優於 FCN-8s 和 DeepLab，加上下文模塊後再提升，在 CamVid、KITTI、Cityscapes 上同樣有效。

街道小圖、藍衣服的例子、「縮小再放大」的格子圖都是示意。第 5 節用一排格子演示感受野 3、7、15、31，原文是二維 3×3 卷積，邊長相同。「認圖部分」指前端，「縮小」指池化和步長。講義未寫具體分數。想深入可看本站的 [第 11 天中文導讀](#)，代碼見開源項目 [30u30 第 11 天目錄](#) (英文)。