The 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)
Taoyuan, Taiwan, October 15-16, 2021. The Association for Computational Linguistics and Chinese Language Processing

RCRNN-based Sound Event Detection System with Specific Speech Resolution
具有特定語音分辨率的 RCRNN 聲音事件偵測系統

黃頌仁 **Sung-Jen Huang,** 王奕雯 **Yih-Wen Wang,** 陳嘉平 **Chia-Ping Chen**
國立中山大學資訊工程學系
National Sun Yat-sen University
Department of Computer Science and Engineering
m093040011@student.nsysu.edu.tw, m083040011@student.nsysu.edu.tw,
cpchen@cse.nsysu.edu.tw
呂仲理 **Chung-Li Lu**, 詹博丞 **Bo-Cheng Chan**
中華電信研究院
Chunghwa Telecom Laboratories
chungli@cht.com.tw, cbc@cht.com.tw

摘要

聲音事件偵測的目標是標記出音訊中的聲音事件及其時間界線。我們基於半監督式學習的均值教師框架，提出一個帶有殘差連接與注意力機制的 RCRNN 網路架構，其可用大量弱標註/未標註資料來訓練。而在許多聲音事件中，語音具有更豐富的訊息量，因此我們使用特定的時間頻率參數來擷取該類別的聲學特徵，並且利用資料增強與後處理來進一步提升效能。我們提出的系統於 DCASE 2021 Task 4 的驗證集上，PSDS (Polyphonic Sound Detection Score)-scenario 1、2 和 Event-based F1-Score 分別達到 38.2%, 58.2% 和 44.3%，優於 baseline 的 33.8%, 52.9% 和 40.7%。

Abstract

Sound event detection (SED) system outputs sound events and their time boundaries in audio signals. We proposed an RCRNN-based SED system with residual connection and convolution block attention mechanism based on the mean-teacher framework of semi-supervised learning. The neural network can be trained with an amount of weakly labeled data and unlabeled data. In addition, we consider that the speech event has more information than other sound events. Thus, we use the specific time-frequency resolution to extract the acoustic feature of the speech event. Furthermore, we apply data augmentation and post-processing to improve the performance. On the DCASE 2021 Task 4 validation set, the proposed system achieves the PSDS (Poly-phonic Sound Event Detection Score)-scenario 1,2 of 38.2%, 58.2% and event-based F1-score of 44.3%, outperforming the baseline score of 33.8%, 52.9% and 40.7%.

關鍵字：聲音事件偵測、均值教師模型、卷積注意力機制、語音

Keywords: Sound event detection, Mean teacher model, CBAM, Speech

1 緒論

聲音在人類的日常生活中無處不在，人們對聲音的接收，是許多行動和反應的判斷基礎，而這通常是基於聲音的事件類別，例如：假若有人呼喊你的名字，你會回頭查看，而若是突然有短促的警報聲，人們則會迅速進入警戒狀態，因此輔助或是代替人們做出決定的決策型機器亦是需要這種判斷聲音事件的能力，由於機器需要能準確的判斷發生的事件類別以及其發生的開始和結束時間，即帶出了聲音事件偵測這個主題，而聲音事件偵測可以簡單的分爲兩類，一類是訓練和測試資料中事件的發生會存在部份重疊，另一類則不會如此，前者在預測上較爲困難，後者則相對容易，DCASE(Detection and Classification of Acoustic Scenes and Events) Challenge Task 4: Sound Event Detection and Separation in Domestic Environments 即是屬於前者，目標

是希望在多個聲音事件彼此重疊的情況下，仍可預測出音訊中所發生的聲音事件及其時間界線。此外，蒐集大量具有時間及事件標註的資料是相當高成本的，因此該任務期望系統可同時利用標註不完全的資料進行訓練。

DCASE 2021 Task 4 的 baseline (Turpault et al., 2019) 是一個基於 CRNN 模型架構的聲音事件偵測系統，且利用均值教師模型 (Tarvainen and Valpola, 2017 ; Lionel and Cyril, 2019) 對弱標註/未標註資料進行半監督式學習。爲了更好的分出十種類別的事件，我們參考 Kim and Kim (2021) 提出的 RCRNN 模型並加以改動，利用由兩層卷積層 (convolution layer) 組合而成的殘差卷積模塊 (residual convolution block) ，目的是希望透過加深層數來增強學習的效果，而其中跳層連接部份的設計則避免了層數過深導致的梯度消失，此外，在十種類別中，我們更加注重語音類別的預測準確度，因此使用了不同解析度的聲學特徵 (Park et al., 2010 ; Zhang et al., 2007)，其中解析度代表的即是透過設定特定的短時傅立葉轉換和梅爾頻率參數來取得與原先不同大小（對特徵的表現也不相同）的梅爾頻譜圖。

論文其餘章節的編排方式將如下安排，章節二：研究方法描述 baseline 系統以及所提出的改進；章節三：實驗設置描述資料集、訊號處理，以及網路參數設定；章節四：實驗結果比較 baseline 系統與改進後系統的效能差異；章節五：結論總結我們系統的優點和未來的研究方向。

2 研究方法

本章節將詳細描述我們提出的聲音事件偵測系統，包含模型架構和半監督式學習的運用，也將說明資料增強的方法以及後處理之具體流程。

2.1 模型架構

2.1.1 CRNN

DCASE Task4 官方提出的 baseline 系統是基於 CRNN 的架構，顧名思義，該架構是由卷積網路 (Convolution Neural Network,CNN) 和遞歸網路 (Recurrent Neural Network,RNN) 組成，其中卷積層可以更好的學習到局部特徵，與之相對的遞歸層則在學習全域特徵上有更佳的表現，而將兩種架構結合後，可同時擁有擷取不同特徵的能力。我們實作了 DCASE Task4 官方提出的 CRNN 系統 (參照圖 1)，此架構使用的是七層卷積層，卷積核大小 (kernel size) 皆是 3×3、激勵函數全部使用門控線性單元 (Gated Linear Unit,GLU)，濾波器 (filter) 數則分別是 16, 32, 64, 128, 128, 128 和 128 個，每層亦使用了批標準化 (batch normalization) 和平均池化 (average pooling)，平均池化的卷積核大小個別是 2×2, 2×2, 1×2, 1×2, 1×2, 1×2 和 1×2，接著連接兩層 128 個單元的雙向門控循環單元 (Bidirectional Gated Recurrent Unit,Bi-GRU) 之遞歸網路架構，最後連接一層乙狀函數 (sigmoid) 的全連接層作爲分類器輸出十個類別的幀級預測 (frame-level prediction or strong prediction)，該預測結果包含事件類別及其時間界線的資訊，接著進一步透過注意力池化層 (Attention pooling) 對幀級預測的時間軸取平均，以得到剪輯級預測 (clip-level prediction or weak prediction)，該預測結果則僅包含事件類別。

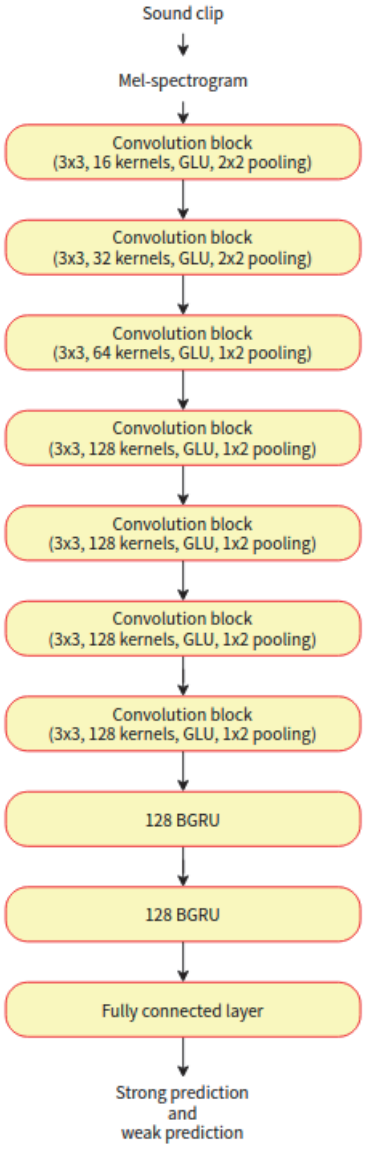

圖 1. Baseline：卷積層中的描述依序代表卷積核大小及濾波器數量、激勵函數和池化層卷積核大小，雙向門控循環單元層的數字則表示單元數。

2.1.2 RCRNN

啓發於 (Kim and Kim, 2021)，我們改良 baseline 模型並實作一個類似的 RCRNN 架構系統，參照圖 2，與其提出的模型相比，我們將殘差連接的部分做了些微改動，此架構將原本 baseline 模型的前兩層卷積核大小改爲 7×7，後五層則將每層改成由兩層卷積模塊和卷積塊的注意力機制 (CBAM,Convolution Block Attention Module) 以及殘差連接組成的殘差模塊（參見圖 2 右半部分），每層卷積模塊使用了與原先大小和數量相同的濾波器（每個殘差模塊視爲一層），同樣使用了批標準化和平均池化，主要不同的是把激勵函數改爲線性整流函數 (Rectified Linear Unit,ReLU)，卷積塊的注意力機制 (Woo et al., 2018) 的運作是如下面公式所示

$$F' = M_c(F) \otimes F \tag{1}$$

$$F'' = M_s(F') \otimes F' \tag{2}$$

$$\begin{aligned} M_c(F) = \sigma(&MLP(Avgpool(F)) \\ &+MLP(Maxpool(F))) \end{aligned} \tag{3}$$

$$M_s(F) = \sigma(f^{(7\times7)}(Avgpool(F); Maxpool(F)) \tag{4}$$

，與常見的注意力模型不同，卷積塊的注意力機制使用乙狀 (sigmoid) 函數（以 σ 表示），而不是歸一化指數函數（softmax），其中 F 爲前一層的輸出，M_c/M_s 是通道/空間注意力機制，$Avgpool/Maxpool$ 是對特徵做平均/最大池化 (即先做通道注意力機制再做空間注意力機制，而通道/空間注意力是使用在空間/通道維度上)，MLP 和 $f^{(7\times7)}$ 分別是僅有一層隱藏層的多層感知器和 7×7 的卷積操作。
由兩層卷積模塊組成是透過加深層數更好的學習十種類別，而加入卷積塊注意力機制的目的則是希望專注於更重要的特徵，最後殘差的設計是避免層數加深導致的梯度消失問題。

2.1.3 半監督式學習

我們參照 baseline 所使用的均值教師框架 (Mean-Teacher framework)，來作爲半監督式學習的方法，一個均值教師模型是由兩個結構完全相同的學生和教師模型組成，只有學生會隨著訓練資料調整其模型的網路參數，教師模型的網路參數則是透過對學生模型的參數進行指數移動平均 (exponential moving average) 後得到，如下式

$$\theta'_t = \alpha\theta'_{t-1} + (1-\alpha)\theta_t \tag{5}$$

，θ 和 θ' 分別表示學生模型和教師模型的參數，t 則代表當前時刻，α 是數值介於 0 和 1 之間的超參數，該框架包含兩種損失函數，分別爲監督式損失 (Supervised loss) 與一致性損失 (Consistency loss)，前者使用二元交叉熵 (Binary Cross-Entropy,BCE)，後者使用平均方差 (Mean Square Error,MSE)。監督式損失用二元交叉熵來計算學生模型對於有標註資料的預測結果與眞實答案的差值，一致性損失則使用平均方差來計算學生模型與教師模型彼此對所有資料預測結果的一致性。此外，一致性損失具有額外的權重，於訓練初期時，將其設定爲零，以便模型優先學習有標註的資料，隨著訓練步數增加，權重亦提高進而開始學習無標註的資料，整體流程如圖 3。

2.2 資料增強

爲了進一步提升效能，我們參照 baseline 所使用的資料混和 (Mixup) 來作爲資料增強的方法 (Zhang et al., 2018)，將兩個資料樣本進行線性組合，以得到新的樣本資料，過程如下

$$\hat{x} = \lambda x_i + (1-\lambda)x_j \tag{6}$$

$$\hat{y} = \lambda y_i + (1-\lambda)y_j \tag{7}$$

，其中 x_i 和 x_j 是隨機選取的兩個樣本的特徵向量，y_i 和 y_j 代表這兩個樣本的標註，$\lambda \in [0,1]$，而其特徵 $\hat{x}$（標註 $\hat{y}$）爲其線性組合產生的新樣本及對應的標註。

2.3 後處理流程

神經網路的幀級預測 (frame-level prediction) 需進一步執行後處理方可得到最終輸出。首先，透過閾值 (Threshold) 將各機率值轉換成二元輸出，接著，再透過中值濾波器 (Median filter) 進一步平滑結果，以避免虛假的預測。我們參照 baseline 所使用的設定，所有事件類別的閾值皆爲 0.5，中值濾波器大小皆爲 7 (即爲 0.45 秒)。

3 實驗設置

3.1 資料集

資料集使用的是由 DCASE 2021 Challenge Task 4 釋出的 DESED(Domestic Environment Sound Event Detection)，當中含有具備兩種資料屬性的三類資料，分別是生成資料：透過 Scaper 工具生成的強標記資料 (標註了音檔中所有發生事件的類別和時間界線) 以及眞實資料：擷取於 Audioset 的弱標記資料 (僅標記發生的事件類別) 和未標記資料，三類資

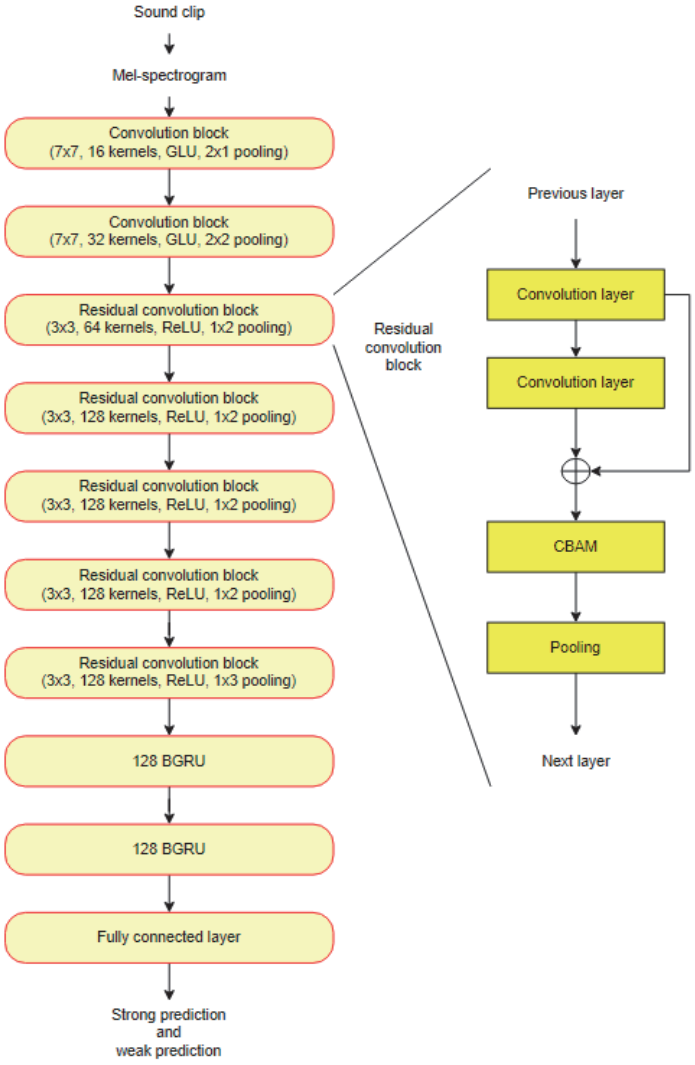

圖 2. 提出的模型：左半部分與圖 1 相似，卷積層和殘差卷積模塊的描述依序爲卷積核大小及濾波器數量、激勵函數和池化層卷積核大小，右半部分則是 residual convolution block 的內部架構

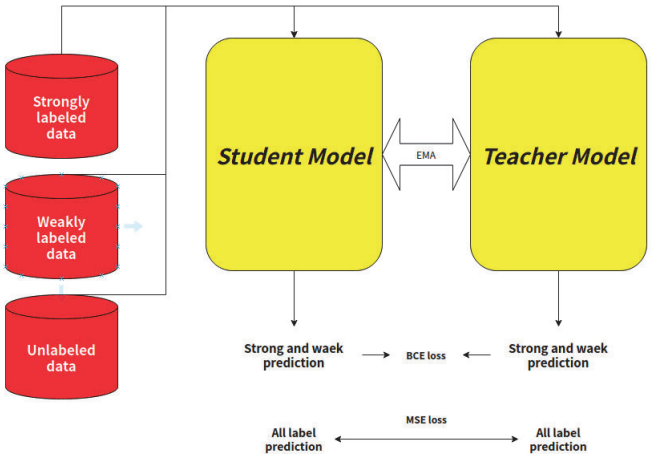

圖 3. 均值教師框架 (Mean-Teacher framework)：將網路架構視爲兩種身分，分別爲學生模型與教師模型。所有資料皆會作爲兩模型的輸入，其中學生模型的參數使用一般梯度更新方式，而教師模型則將學生模型參數進行指數移動平均，以得到當前時刻的參數。此流程包含兩種損失函數，分別爲監督式損失與一致性損失。

	Baseline	RCRNN(A)	RCRNN(B)
Alarm/bell/ringing	42.5%	44.7%	**46.0%**
Blender	46.8%	37.9%	**48.7%**
Cat	42.4%	46.6%	**48.2%**
Dishes	22.7%	32.2%	**33.4%**
Dog	25.8%	**27.1%**	**27.1%**
Electric shaver/toothbrush	45.4%	**59.0%**	51.0%
Frying	34.3%	**42.6%**	37.2%
Running water	37.8%	39.4%	**43.8%**
Speech	53.0%	53.6%	**61.4%**
Vacuum cleaner	**56.1%**	47.7%	46.3%

表 1. 各類別比較：在不同解析度與不同網路架構下，各類別的 Event-based F1 (event-f1) Score

料個別有 10000、1578 和 14412 筆，驗證資料和測試資料分別是 2500 和 1168 筆強預測資料，詳細資訊如表 2 所示。

	訓練資料			驗證資料	測試資料
	強標記	弱標記	未標記	強標記	強標記
數量	10000	1578	14412	2500	1168
種類	生成	眞實	眞實	生成	生成
長度	至多 10 秒				
通道數	1	2	2	1	1
採樣率 (kHz)	16	44.1	44.1	16	16

表 2. DESED 資料集

3.2 訊號處理

在 DESED 資料中，所有音檔的採樣率與聲道數並非一致，因此我們先利用 FFmpeg 工具將所有資料屬性統一爲 16000 Hz 和單聲道，接著透過 Librosa 工具從音檔中擷取聲學特徵來做爲神經網路的輸入，而這裡擷取的聲學特徵是梅爾頻譜圖。

3.3 訓練設定

我們提出的系統使用的皆是 RCRNN 架構，下面會分爲是否改動解析度的兩類描述，如此改動是爲了提升語音類別的準確度，主要差異爲梅爾頻譜圖參數設置及池化層稍有改動，而 baseline 的設置則參照章節 2.1.1 描述（解析度與未改動的 RCRNN 模型相同）。

3.3.1 未改動解析度的 RCRNN 模型

參數設置如圖 2 左半部分所示，生成梅爾頻譜圖的參數 (版本 A) 爲 n_mels: 128 (128 個 mel-filter bank)、n_filters: 2048 (離散傅立葉轉換的樣本數)、hop_length: 256 (短時傅立葉轉換的 window 間隔)、n_window: 2048 (短時傅立葉轉換的 window 大小)。

3.3.2 改變解析度的 RCRNN

爲了能在語音類別上有更佳的準確度，將聲學特徵擷取參數改變以擷取到更爲清晰的語音類別特徵，因此在採樣頻率保持 16000 Hz 下將生成梅爾頻譜圖的參數 (版本 B) 改爲 n_mels: 96、n_filters: 2048 、hop_length: 192 、n_window: 1536（參考表 3），以上的單位除了 mel-filter bank 外皆是樣本數，另外爲了維持最終輸出大小與擷取出的標註大小相同，將第一層和最後一層池化層大小改爲 2×1 和 1×3，如此的設計是由於在 (Benito-Gorrón et al., 2021) 的實驗中 PSDS 近似的情況下，語音類別的準確度有相當的提昇，其中的研究也指出不同類別事件在不同聲學特徵下有不同表現（例如電動刮鬍刀的聲音類別在頻率擷取更密集的聲學特徵下有更清晰的表現，警示聲則在時間擷取更密集的特徵下更明顯）。

	A	B
n_mels	128	96
n_filters	2048	2048
hop_length	256	192
n_window	2048	1536

表 3. 聲學特徵參數設定：爲了提升語音類別準確度，將聲學參數由版本 A 改爲版本 B

	PSDS-1	PSDS-2	event-f1
Baseline	0.338	0.529	40.7%
RCRNN	**0.374**	**0.563**	**43.1%**

表 4. 架構改動結果：在相同解析度下，不同網路架構的實驗結果

	解析度	PSDS-1	PSDS-2	event-f1
Baseline	A	0.338	0.529	40.7%
RCRNN	A	0.374	0.563	43.1%
	B	**0.382**	**0.582**	**44.3%**

表 5. 解析度改動結果：在不同解析度與不同網路架構下的實驗結果

4 實驗結果

4.1 模型架構改動結果

表 4 呈現 baseline 系統與我們所提出的系統之效能比較。在各評估標準下，RCRNN 模型皆明顯優於 CRNN 模型，於 PSDS scenario 1 由 0.338 提升至 0.374，PSDS scenario 2 亦由 0.529 提升至 0.563，event-f1 則由 40.7% 提升至 43.1%，顯示層數加深並且使用卷積注意力機制對於效果有顯著提昇。

4.2 解析度改動結果

表 5 列出在不同解析度設定的情況下，不同網路架構的實驗結果，而在表 1. 各類別比較中顯示了在 CRNN 模型和兩種解析度的 RCRNN 模型中各個類別事件預測的 f1-score，從表 1、4 和 5 的結果可以看出 RCRNN 模型在提升所有分數的同時，語音類別也能有更高的正確性，而改動解析度的模型除了所有分數再次提昇外，語音類別的分數亦大幅提昇，令語音類別 f1-score 達到 61.4%，可見在稍微增加時間解析度且少量減少頻率解析度下對於識別語音類別是有明顯幫助的。

5 結論

我們提出了一個運用不同聲學參數梅爾頻譜圖的 RCRNN 架構系統，透過模型架構的改動提升了整體的分數，而改變解析度極大的提升語音類別的準確度，同時，整體分數（PSDS、event f1-score）也有小幅提昇，觀察 baseline 系統和 RCRNN 系統的比較表可見，RCRNN 是明顯優於 baseline 系統的，使用不同的解析度後，語音類別的預測有明顯的提昇，而未來研究的方向大致有兩個面向，其一是研究 RCRNN 模型架構下適合何種參數設置（例如 CNN 的卷積核大小和濾波器數量），二則是研究何種解析度對於語音類別是最佳的，以及探討此種解析度會更好的原因。

References

Diego De Benito-Gorrón, Daniel Ramos, and Doroteo T. Toledano. 2021. A multi-resolution crnn-based approach for semi-supervised sound event detection in dcase 2020 challenge. *IEEE Access*, 9:89029–89042.

Nam Kyun Kim and Hong Kook Kim. 2021. Polyphonic sound event detection based on residual convolutional recurrent neural network with semi-supervised loss function. *IEEE Access*, 9:7564–7575.

Delphin-Poulat Lionel and Plapous Cyril. 2019. Mean teacher with data augmentation for dcase 2019 task 4. In *Detection and Classification of Acoustic Scenes and Events 2019*.

Dennis Park, Deva Ramanan, and Charless Fowlkes. 2010. Multiresolution models for object detection. In *European Conference on Computer Vision(ECCV) 2010*, pages 241–254.

Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*.

Nicolas Turpault, Romain Serizel, Ankit Parag Shah, and Justin Salamon. 2019. Sound event detection in domestic environments with weakly labeled data and soundscape synthesis. In *Detection and Classification of Acoustic Scenes and Events*.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV), 2018*, pages 3–19.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations(ICLR 2018)*.

Wei Zhang, Gregory Zelinsky, and Dimitris Samaras. 2007. Real-time accurate object detection using multiple resolutions. In *2007 IEEE 11th International Conference on Computer Vision*.