人工智慧概論 期末專題

姓名：杜承祐

學號：4111056009

1. **研究背景**

隨著人工智慧與深度學習技術的發展，影像辨識已廣泛應用於各種場景，其中情緒辨識更是與人類日常生活密切相關的重要議題。透過分析人臉表情來判斷情緒狀態，不僅能應用在智慧教學、互動機器人等領域，也有助於實現更人性化的智慧系統。

目前已有多個公開的情緒資料集與預訓練模型可供使用，使得我們能以遷移學習方式進行表情分類訓練，並觀察不同模型與資料集對準確度與泛化能力的影響。

1. **研究動機**

在過往的專題或課程中，我們對卷積神經網路已有基本認識，然而對於「不同模型架構在實際任務中的表現差異」仍缺乏深入了解。同時，許多現有的情緒資料集都集中在基本情緒，而對於像「學習情緒」這種更抽象的分類，相關研究與資料較少。

因此，我們希望透過本次作業，實際操作多種模型與資料集的訓練流程，並進一步延伸到自建資料集與進階情緒分類任務，以增進我們對影像分類與遷移學習的實務理解。

1. **目的**

本次實驗主要目標為：

1. 比較三種不同影像分類模型（VGG19、ResNet50、EfficientNetB0）在不同公開情緒資料集上的訓練表現，了解其特性與差異。
2. 建立自建表情資料集，並進行遷移學習測試，驗證模型是否能辨識實際個體的情緒。
3. 進一步延伸訓練內容，將基本情緒辨識模型遷移到「學習情緒」這類抽象情緒資料集，評估模型的泛化與應用潛力。
4. 綜合分析不同模型與資料集的準確率、混淆矩陣、錯誤案例與特徵可視化結果，作為未來應用與改進的參考。
5. **研究方法**

本研究旨在探討不同深度學習模型於不同情緒資料集上的表現差異，透過遷移學習方法進行情緒辨識任務。整體實驗分為三個主要部分：模型選擇、資料集處理與訓練流程設計，詳細說明如下：

### **4.1 模型選擇**

本實驗選用三種具有代表性的深度卷積神經網路架構，分別為：

1. VGG19：具備簡單層疊結構與較深的層數，常作為基準模型。
2. ResNet50：引入殘差連接（residual connections）以解決深層網路的退化問題，提升訓練穩定性。
3. EfficientNetB0：透過複合比例縮放（compound scaling）策略，在保持高準確率的同時有效減少參數量與計算成本。

### **4.2 資料集設定**

選擇之公開資料集分別為REF2013 [1], REFplus [2], RAF-DB [3]，為進行全面比較，共產生 3（模型） × 3（資料集） = 9 組模型訓練實驗。每個模型僅針對單一資料集訓練，不進行跨資料集預訓練，以便觀察模型對資料集特性之適應能力，訓練資料分割依照80%,20%進行。

### **4.3 訓練流程設計**

每一組模型訓練過程分為三個階段的微調策略，大致流程如下：

#### 特徵提取：將預訓練模型最深層的卷積區塊解凍，同時訓練新增的全連接層（Dense layers）以適應新資料集分類任務。

#### 部分解凍微調：解凍模型中更多的卷積區塊，搭配較低學習率進行訓練，增強中高層特徵對目標任務的辨識能力。

#### 最終模型微調：解凍到大約中層卷積區塊，並以較低學習率進行微調，使模型整體參數能根據資料集特性進行最終調整。

此階段訓練超參數與解凍層數等視各模型在驗證集的收斂情況彈性調整。

### **4.4 訓練與測試設定**

1. 輸入尺寸：皆調整至224x224，統一輸入維度。
2. 資料增強（Augmentation）：包含隨機旋轉、水平翻轉、縮放等方式，提升模型泛化能力。
3. 損失函數：categorical crossentropy。
4. 優化器：Adam（含自動調整學習率機制）。
5. 評估指標：主要以測試集的分類準確率（accuracy）為依據，並搭配混淆矩陣與 Grad-CAM 可視化進行進一步分析。

#### **4.5 自製臉部影像資料集實驗**

#### 為評估所訓練模型對真實應用場景之適應能力，本研究拍攝並建立個人臉部情緒影像小型資料集，涵蓋基本情緒類別（如：開心、難過、生氣、驚訝、中立），並由本人進行人工標記，將原本已訓練完成之 VGG19/ResNet/EfficientNet 模型以小型自製資料集進行遷移微調，或直接以此資料輸入原模型測試其泛化能力。

**4.6 學習情緒辨識任務**

為突破基本情緒分類限制，本研究採用Student Engagement [4]進行第二項延伸：將基本情緒模型遷移至學習情緒分類任務。學習情緒（如：專注、疲倦、焦慮、自信）雖較抽象，但與實際應用（如智慧教室、遠距教學）密切相關，因此嘗試以下流程：

1. 小型學習情緒資料集建立：參考公開學習情緒資料集。
2. 遷移學習流程：採用本研究訓練完成的 VGG19、ResNet50 或 EfficientNetB0 基本情緒模型作為起點，調整分類層數量，進行短期微調以適應新任務。
3. 效果分析：評估在學習情緒資料上的辨識準確率，並與原始基本情緒任務進行比較，探討學習情緒的分類挑戰性與模型遷移性。

**5. 實驗結果**

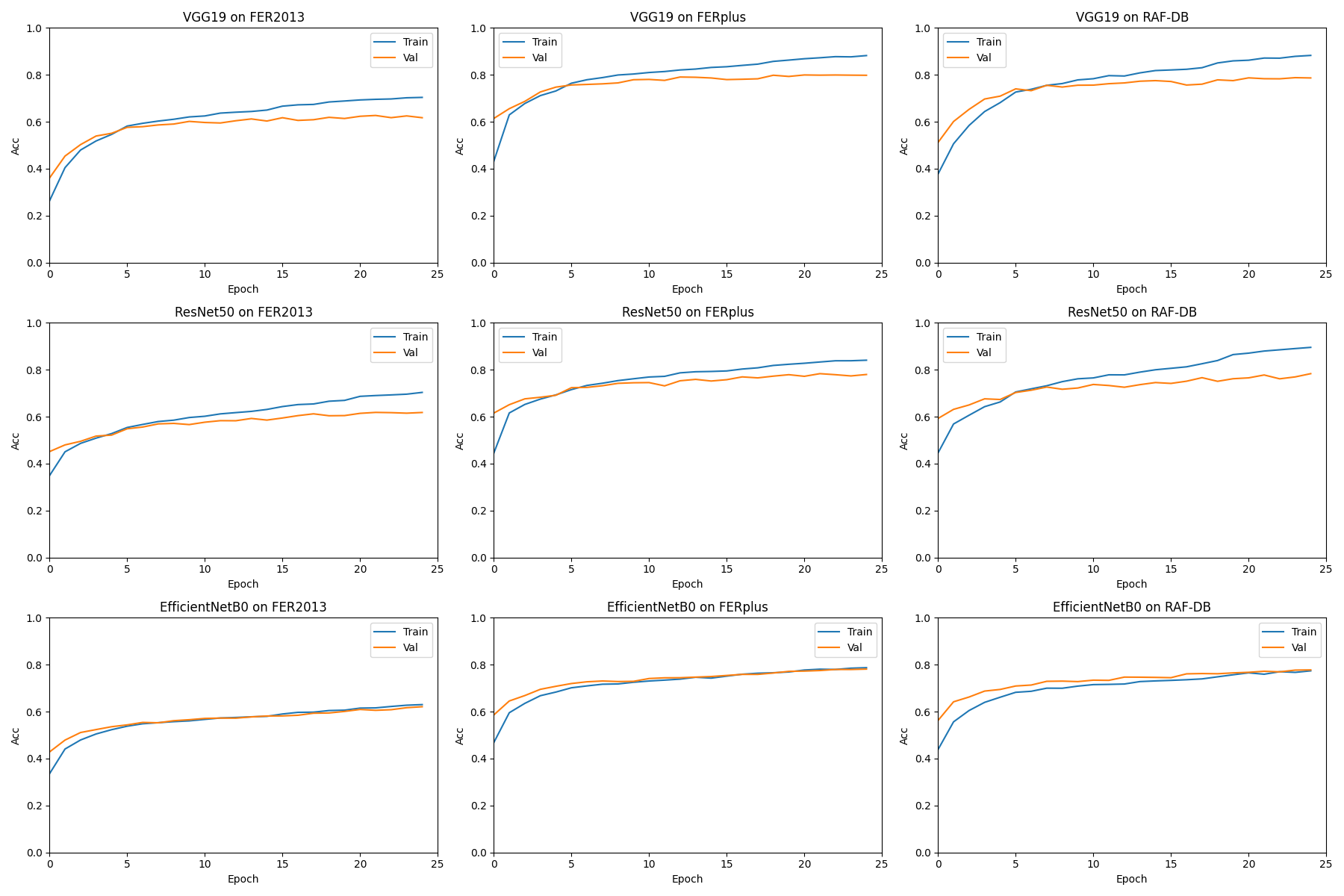
**5.1 基本情緒遷移訓練**

此部分針對三種模型與三種資料集之遷移訓練實驗結果進行分析與比較。

1. 訓練與驗證之準確率學習曲線：

從橫向比較來看，即相同模型在不同資料集下的訓練表現，可以發現無論使用哪種模型，在 FER2013 資料集上的表現普遍較差；而表現最好的則是 FERplus，其次為 RAF-DB。這顯示 FER2013 的原始標註品質較差，經過更完善的重新標註後，FERplus 的訓練效果有明顯提升。至於 RAF-DB，雖然標註品質不錯，但由於資料量相對較小，可能對模型的學習仍有些微限制。

從縱向比較來看，也就是不同模型在相同資料集上的表現，可以觀察到 VGG19 在訓練與驗證準確率（train acc 與 val acc）之間有明顯落差，顯示有過擬合的傾向；相對地，EfficientNetB0 的兩者表現較為接近，表示其具有較好的抗過擬合能力與泛化性。而 ResNet50 則在 RAF-DB 上出現明顯的過擬合，顯示這類模型更適合用於如 ImageNet 等大型資料集，當資料量不足時，其深層結構反而可能無法發揮應有的效果。

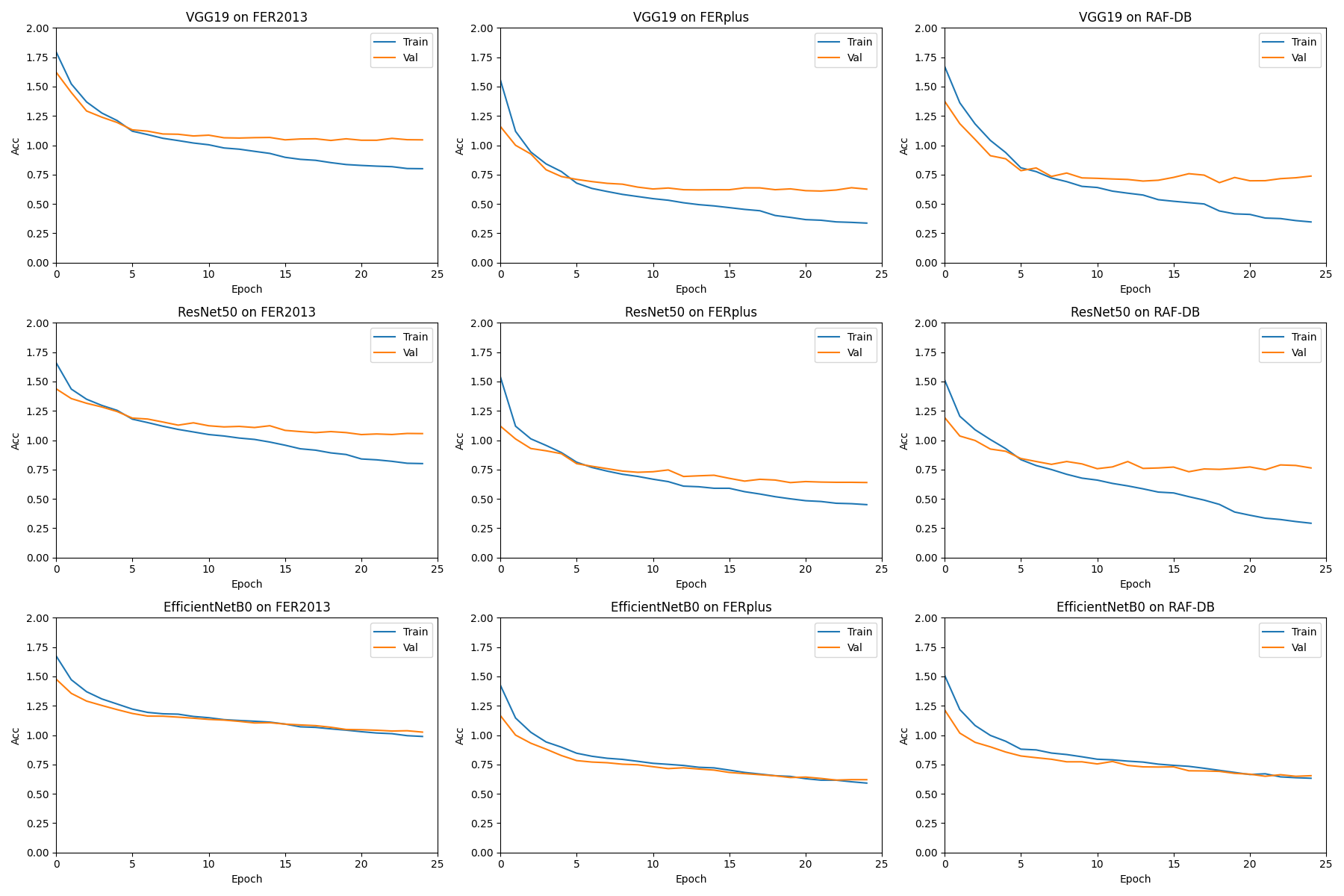


(圖一) 各實驗之訓練與驗證準確率曲線

1. 訓練與驗證之損失函數學習曲線：

從橫向比較來看，也可以觀察到類似的趨勢：在 FER2013 資料集上，損失函數值整體下降幅度有限，顯示模型在該資料集上的學習效果不佳。

進一步從不同模型的角度分析，可以發現 VGG19 與 ResNet50 在訓練中後期雖然 train loss 持續緩慢下降，但 val loss 卻未見明顯改善，表示泛化能力在這個階段已經無法再提升，甚至可能出現過擬合現象。這種情況在 RAF-DB 資料集中尤其明顯，可能與該資料集規模較小有關。相對地，EfficientNetB0 的損失曲線則呈現訓練與驗證之間的緊密貼合，顯示該模型在學習過程中未過度擬合訓練資料中的細節，維持良好的泛化效果，也說明其較強的抗過擬合能力。



(圖二) 各實驗之訓練與驗證損失函數曲線

1. 訓練、驗證與測試準確率

首先從不同模型在三個資料集的表現來看，VGG19 的驗證準確率與測試準確率在各資料集中差異不大，代表模型在訓練完成後能穩定地在測試資料上展現不錯的效果；而在 ResNet50 的實驗中，可以明顯觀察到 RAF-DB 測試準確率明顯低於驗證準確率，且與訓練準確率之間差距更大，顯示模型在該資料集上已出現過擬合的現象；至於 EfficientNetB0，三組實驗中測試準確率普遍低於驗證準確率，顯示雖然訓練過程穩定，但模型的學習速度較慢，在相同訓練輪次下，無法充分學習資料特徵，可能處於輕微的欠擬合狀態。

進一步從資料集角度分析，不同模型在 FERplus 上普遍表現最好，而與 FER2013 相比常有超過 15% 的差距，顯示更高品質的標註資料確實能有效提升模型效能。特別的是，在 EfficientNetB0 的結果中，RAF-DB 的表現甚至略優於 FERplus，可以推測對於學習速度較慢的模型來說，使用較小、較乾淨的資料集反而能更快學習到有效特徵。

(表一) 各實驗之訓練、驗證與測試準確率

| Model | Dataset | Train Accuracy | Validation Accuracy | Test Accuracy |
| --- | --- | --- | --- | --- |
| VGG19 | FER2013 | 70.38% | 61.73% | 62.97% |
| VGG19 | FERplus | 88.20% | 79.80% | 80.72% |
| VGG19 | RAF-DB | 88.27% | 78.67% | 78.36% |
| ResNet50 | FER2013 | 70.36% | 61.80% | 62.65% |
| ResNet50 | FERplus | 84.07% | 77.99% | 78.86% |
| ResNet50 | RAF-DB | 89.54% | 78.34% | 75.49% |
| EfficientNetB0 | FER2013 | 63.00% | 62.10% | 59.43% |
| EfficientNetB0 | FERplus | 78.72% | 78.13% | 74.88% |
| EfficientNetB0 | RAF-DB | 77.38% | 77.77% | 75.65% |

1. precision, recall, F1-score

觀察不同組實驗的三種分類指標（Precision、Recall、F1-score），分別以 macro 與 weighted 方式進行評估，可以發現大多數情況下，weighted 分數高於 macro 分數，顯示模型傾向學習樣本數較多的類別，而對樣本較少的類別則學習效果有限。其中在 EfficientNetB0 on FERplus 這組實驗中，F1-score 的 macro 與 weighted 差距高達 25%，顯示類別不平衡對模型表現影響極大。

唯一出現 macro precision 高於 weighted precision 的情況是 EfficientNetB0 on FER2013，這可能代表模型對少數類別的預測數雖少，但準確率高（保守但精確），因此 macro precision 被拉高；然而因預測量少，Recall 難以提升，使得 F1-score 整體拉不上來。這種現象可能來自於資料集的特性：樣本數較多的類別含有較多錯誤標註，導致學習混亂，而樣本數少的類別反而因特徵明顯、標註相對準確，導致 precision 偏高。

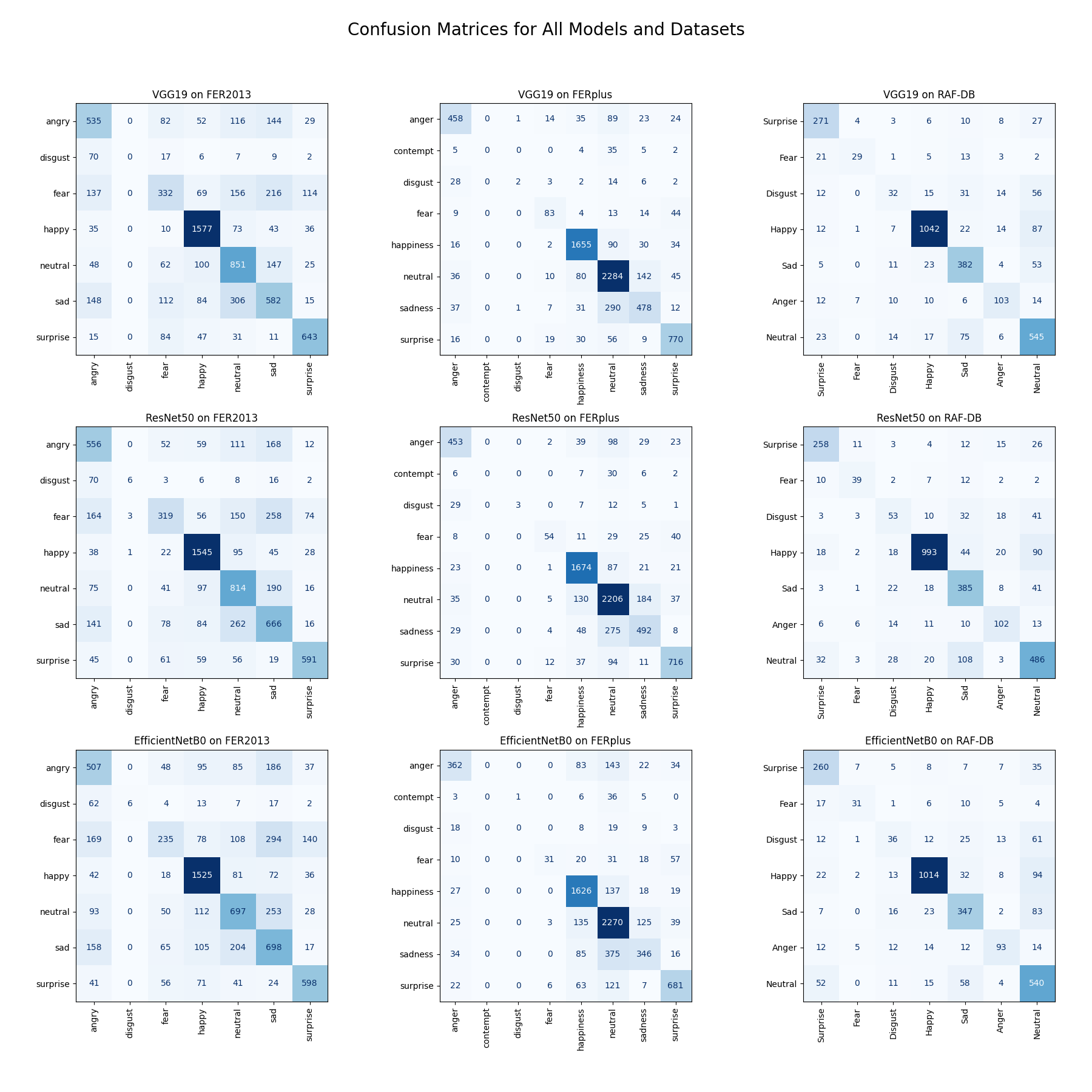
(表二) 各實驗之precision, recall, F1-score

| Model | Dataset | Precision Macro | Precision Weighted | Recall Macro | Recall Weighted | F1 Macro | F1 Weighted |
| --- | --- | --- | --- | --- | --- | --- | --- |
| VGG19 | FER2013 | 51.91% | 61.03% | 52.89% | 62.97% | 52.05% | 61.61% |
| VGG19 | FERplus | 63.18% | 79.54% | 55.53% | 80.72% | 56.66% | 79.80% |
| VGG19 | RAF-DB | 69.89% | 78.04% | 64.73% | 78.36% | 66.02% | 77.73% |
| ResNet50 | FER2013 | 61.54% | 62.78% | 53.18% | 62.65% | 53.44% | 61.68% |
| ResNet50 | FERplus | 69.36% | 78.08% | 52.69% | 78.86% | 54.79% | 77.86% |
| ResNet50 | RAF-DB | 66.22% | 76.45% | 66.15% | 75.49% | 65.95% | 75.70% |
| EfficientNetB0 | FER2013 | 63.55% | 59.47% | 50.24% | 59.43% | 50.00% | 57.91% |
| EfficientNetB0 | FERplus | 55.70% | 73.30% | 45.91% | 74.88% | 47.99% | 72.96% |
| EfficientNetB0 | RAF-DB | 67.53% | 75.92% | 62.63% | 75.65% | 64.08% | 75.31% |

1. 混淆矩陣

綜觀三個資料集中的類別分布，可以發現都存在某些類別佔據多數樣本的情況，因此模型只要偏向預測這些主要類別，就能取得不錯的整體準確率。在 FER2013 中，*disgust* 類別甚至出現「預測正確數為 0」的情況，且大多被錯誤判斷為 *angry*，顯示兩者在表情上較為相似，而 *disgust* 樣本數又極少，導致模型難以學習該類別特徵。

在 FERplus 的實驗中也有類似現象，*contempt* 類別大量被預測為 *neutral*，說明此類表情不夠明顯，模型難以分辨其與其他類別間的細微差異。這些結果顯示：資料數量不足的類別若同時缺乏獨特的特徵，容易被其他相似類別覆蓋，造成模型無法有效區分。



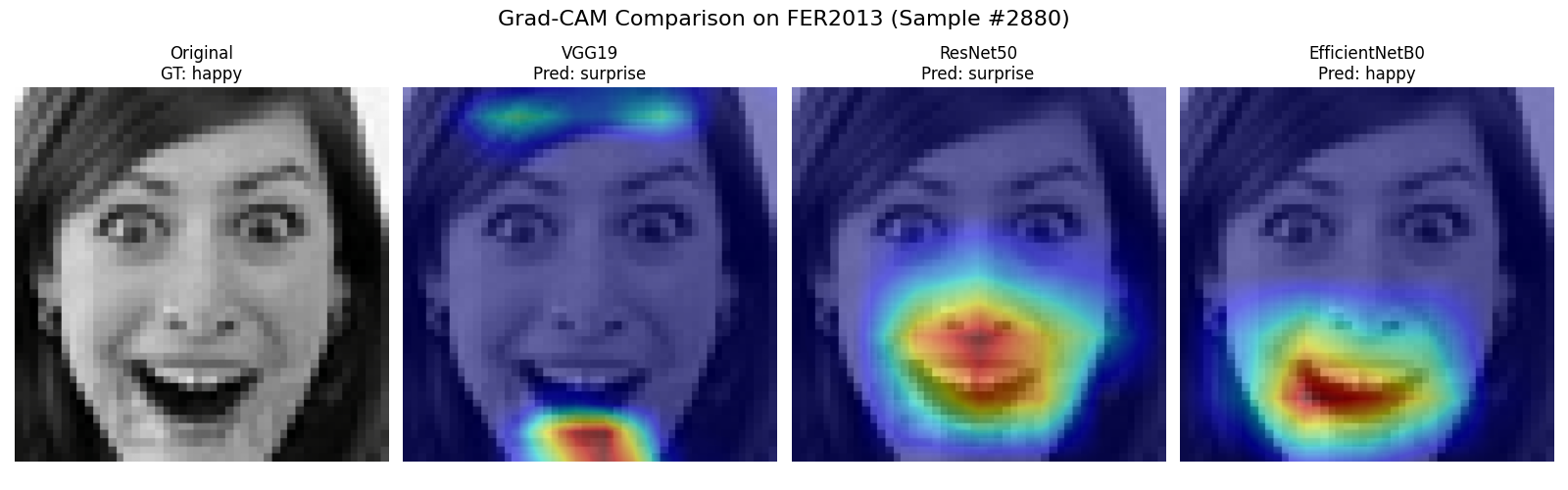
(圖三) 各實驗之混淆矩陣圖

1. GradCam與錯誤案例分析

本部分針對三個資料集，各隨機選取一張測試集圖片進行 Grad-CAM 視覺化，觀察不同模型所聚焦的區域，以及當預測錯誤時的差異。Grad-CAM 所追蹤的卷積層皆為各預訓練模型的最後一層，輸出特徵圖大小為 7×7，因此與原圖重疊時可能出現位置偏差。

* FER2013 - happy

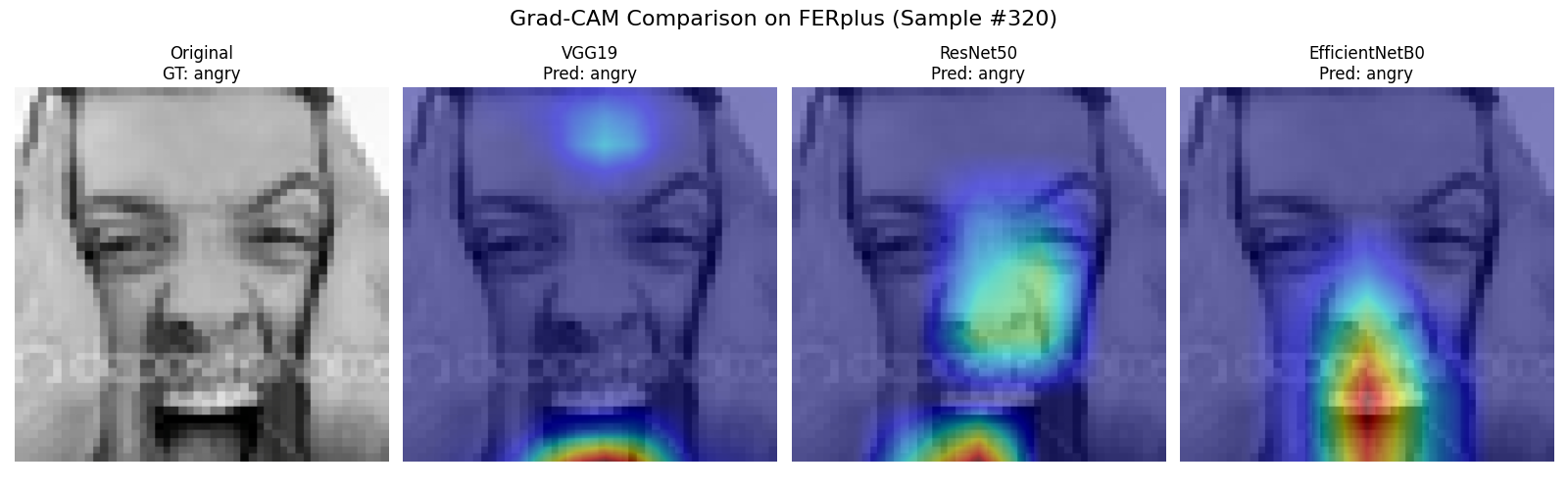
僅有 EfficientNetB0 成功預測為 *happy*，其他模型皆誤判為 *surprise*。三個模型都較偏向依賴嘴型作為主要判斷依據，其中 VGG19 的 Grad-CAM 還能觀察到其對眼睛的注意力。然而，開心與驚訝的嘴型都可能較為寬開，且眼睛睜大的特徵在驚訝中更為常見，因此容易出現誤判。



(圖四) FER2013 - happy 各模型預測結果與GradCam可視化

* FERplus - angry

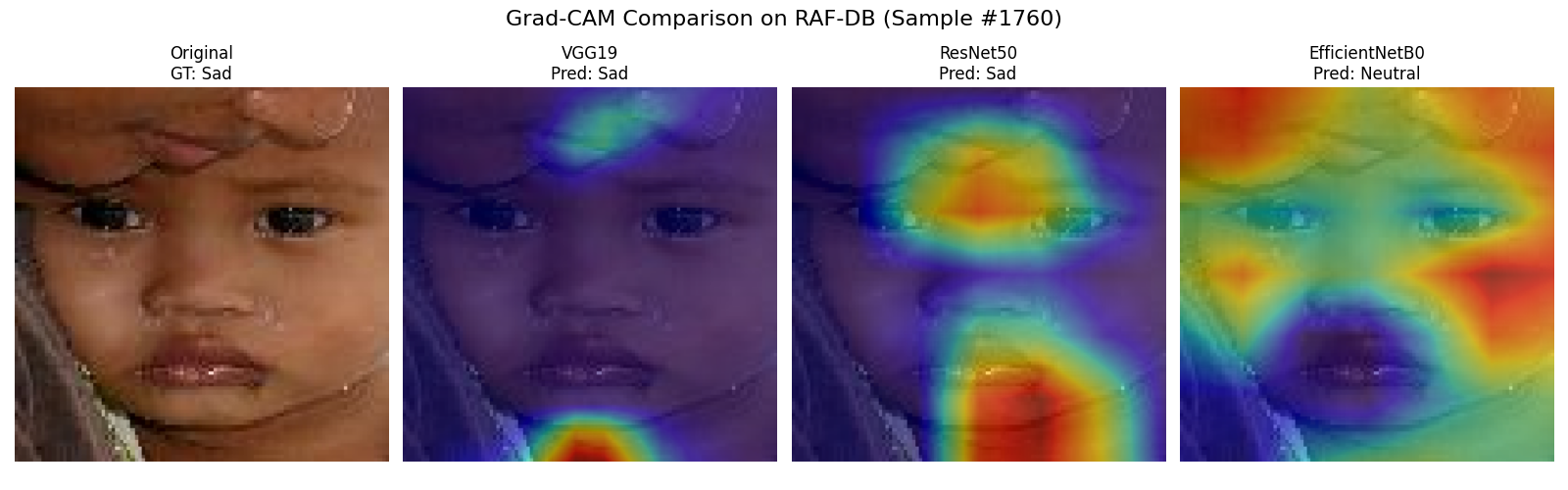
三個模型皆成功預測為 *angry*，Grad-CAM 顯示模型普遍關注嘴型，部分也注意到眉心附近的區域。這些特徵符合憤怒表情的特性，例如拉長的嘴型與緊皺的眉毛，說明模型能夠正確捕捉並利用情緒表徵。



(圖五) FERplus - angry 各模型預測結果與GradCam可視化

* RAF-DB - sad

本案例中僅有 EfficientNetB0 預測錯誤。從 Grad-CAM 可見，EfficientNetB0 沒有關注嘴型，而其他兩個模型除了嘴部外，還聚焦於眉毛附近。傷心表情通常表現為嘴角下垂與眉毛內縮、下垂等特徵，顯示前兩者成功利用這些關鍵部位進行正確預測，而 EfficientNetB0 缺乏對這些區域的注意，導致誤判。



(圖六) RAF-DB - sad 各模型預測結果與GradCam可視化

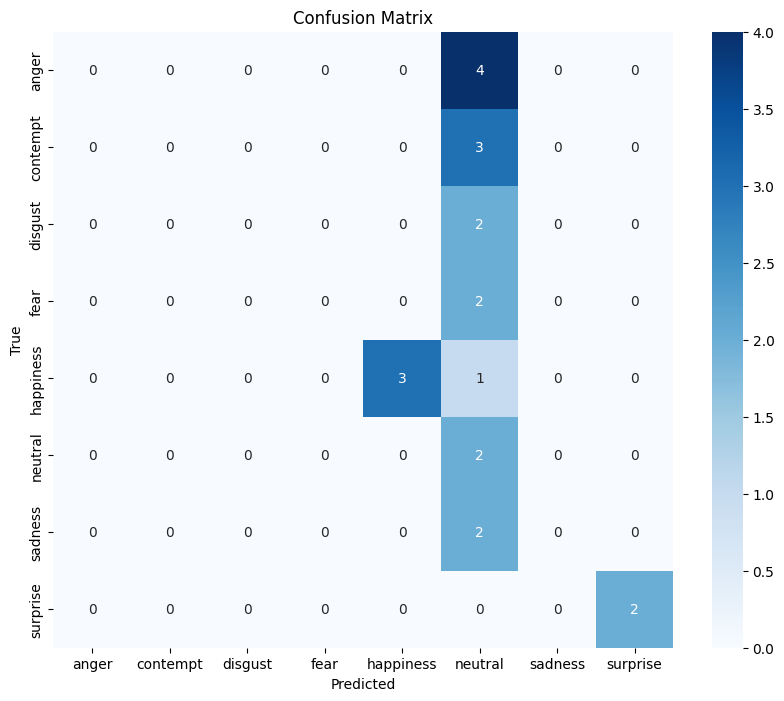
**5.2 自建資料集之模型效能測試**

本步驟中，我們拍攝自身的表情並收集不同情緒類型，以建立自建資料集。測試時使用的模型為在基礎遷移訓練實驗中表現最穩定之 **VGG19 on FERplus**，情緒類別亦參照 FERplus 所定義的八種類別。

1. 測試結果

* Test Accuracy: 0.3333
* Test Loss: 7.9682

可以發現整體測試表現不理想，僅在 **happiness** 與 **surprise** 這兩類表情特徵明顯的類別上出現正確預測。其餘類別由於表情不夠清楚或表現不精確，導致模型大多將其誤判為 **neutral**。



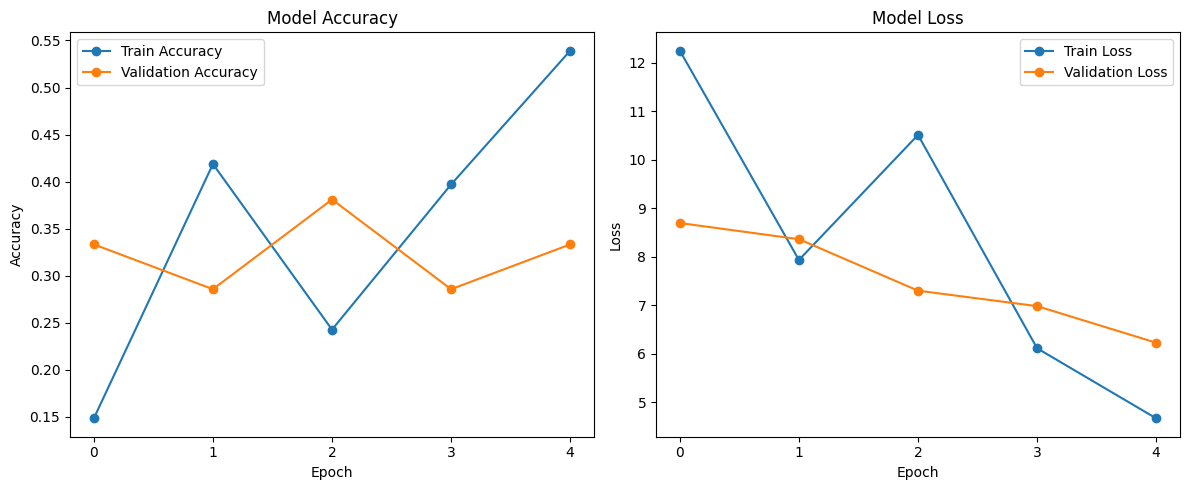
(圖七) 自建資料集之測試混淆矩陣

1. 遷移訓練

由於資料集規模較小，本次訓練僅微調分類頭，進行 **5 輪**訓練，並採用小學習率與適度的資料增強，以避免模型僅記住資料本身而非學習有效特徵。

* Train Accuracy: 0.5391
* Val Accuracy: 0.3333

實驗結果顯示，訓練準確率雖有一定提升，訓練與驗證損失亦有些微下降，然而驗證準確率卻未見顯著改善。這顯示資料量不足，導致模型難以有效捕捉類別內共通特徵，進而無法辨識不同表情的細節差異。

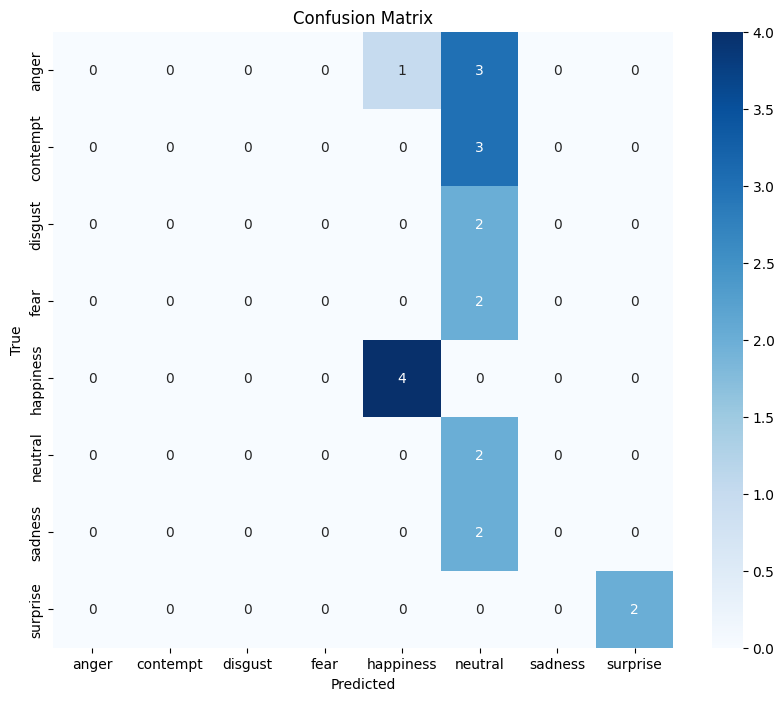


(圖八) 自建資料集之遷移訓練與損失曲線

1. 遷移訓練後之測試結果

* Test Accuracy: 0.3810
* Test Loss: 6.7512

從測試結果可見，準確率與損失雖有些微改善，但僅 **happiness** 類別的預測效果有所提升。整體而言，模型仍主要集中於特徵明顯或出現頻率較高的類別，對於定義模糊的情緒類別仍難以準確分類。若欲進一步提升模型在自建資料集下的辨識能力，必須擴大資料集數量與多樣性，並解凍部分卷積層進行進一步微調，以在不造成過擬合的前提下，實現更有效的遷移學習。



(圖九) 自建資料集遷移學習後之測試混淆矩陣

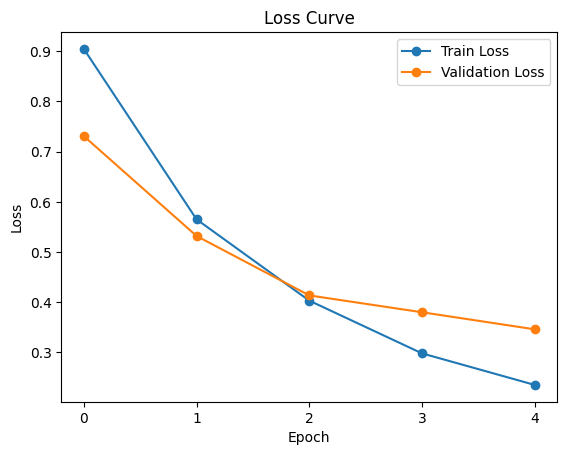
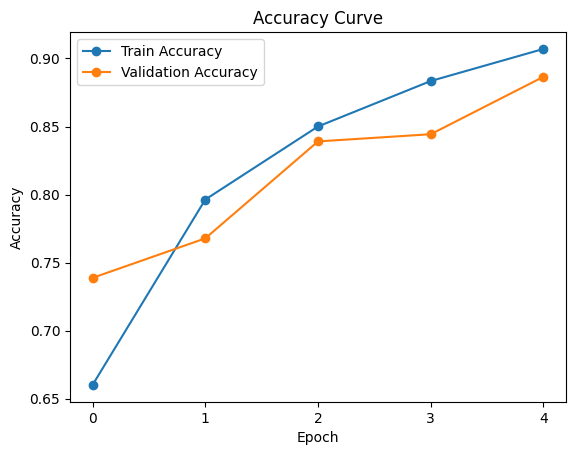
**5.3 學習情緒遷移訓練**

在本階段中，我們使用一個包含六種學習過程中常見表情的「學習情緒」資料集進行遷移訓練。由於資料集規模較小，我們基於基礎情緒分類之實驗結果，選擇以於FERplus上預訓練的 VGG19 模型作為基礎，以確保較佳的泛化能力與分類效能。

資料集：Student-engagement [4]

1. 訓練與驗證之Accuracy與loss曲線

從圖中可觀察到，模型在 10 輪訓練內即能快速提升準確率並降低損失值，顯示學習穩定且有效。



(圖十) 訓練與驗證之Accuracy與loss曲線

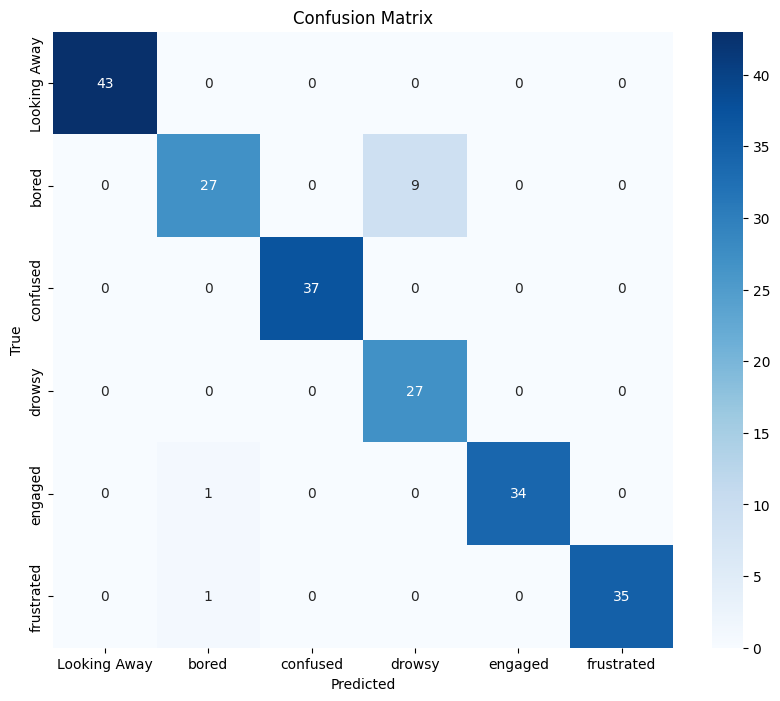
1. 準確率

* 訓練準確率：90.29%
* 驗證準確率：88.65%
* 測試準確率：94.86%

可以發現測試與驗證的準確率皆與訓練準確率相當，顯示模型並未過擬合，且具有良好的泛化能力。

1. 混淆矩陣

大部分類別皆能被模型有效辨識，僅在 *bored* 與 *drowsy* 類別上出現部分混淆。可能原因在於兩者皆可能出現眼睛半閉等類似特徵，導致模型無法清楚區分。



(圖十一) 學習情緒之混淆矩陣

1. 錯誤案例

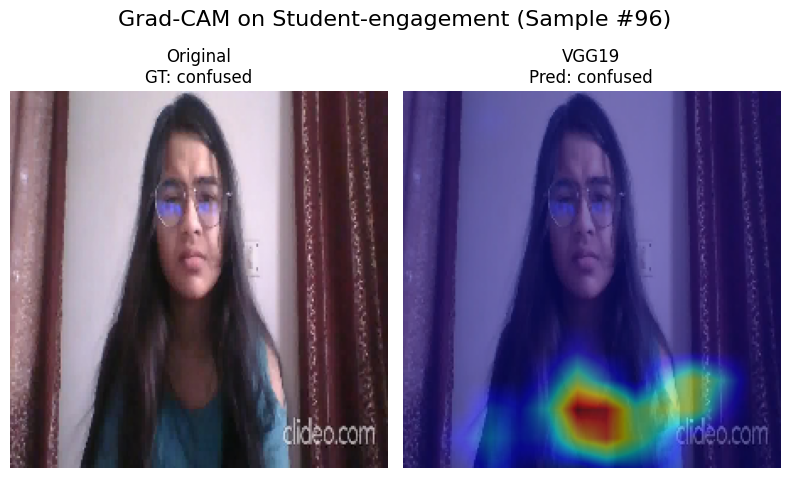
於此案例中，標註標籤為 *bored*，但影像中學生頭部明顯下垂，極可能處於打瞌睡狀態，導致模型將其誤判為 *drowsy*。此例顯示資料標註與實際特徵存在一定模糊性。



(圖十二) 錯誤案例示意圖

1. GradCam

儘管此測試樣本的預測結果為正確，但 Grad-CAM 可視化結果顯示模型專注的區域並非臉部特徵，而是學生的衣著。這可能是由於資料集中部分類別的樣本數偏少，且來自單一學生，導致模型學習到以衣著作為判斷依據的偏差特徵。此結果顯示資料集仍存在侷限性，未來需擴充資料量並提升多樣性，以促進模型學習真實且具代表性的表情特徵。



(圖十三) GradCam視覺化

**6. 總結**

本次實驗透過比較三種常見的影像特徵擷取模型與三個具代表性的人臉表情分類資料集，進行表情分類的遷移訓練，並觀察各種模型與資料組合的訓練特性與表現差異。

在模型方面，我們發現 VGG19 在中小型任務中能夠快速收斂並達成良好效果；EfficientNetB0 雖然透過多種優化策略成功減少參數量，卻在本任務中出現 欠擬合 的情形；而 ResNet50 則在效能與穩定性之間呈現相對平衡的表現。

資料集方面，FER2013 與 FERplus 的訓練結果對比顯示，資料標註品質對模型準確率有顯著影響，差距甚至超過 15%。至於 RAF-DB 這類規模較小的資料集，則較容易在相同訓練輪次下出現 過擬合 現象，需特別注意訓練策略與監控指標。

針對自建資料集的遷移訓練，我們觀察到在資料數量非常有限的情況下，模型容易出現過擬合；即使採取資料增強等改善策略，也難以讓模型有效學習並遷移到新資料，顯示若欲提升辨識能力，仍需蒐集更多樣且充足的資料並進一步優化訓練流程。

最後，在嘗試將模型應用於六種抽象情緒的學習任務中，發現若使用以基本情緒訓練出的模型進行遷移學習，仍具備一定程度的辨識能力。但資料集本身在部分情緒類別上存在 樣本不足與標註品質不一 等問題，導致整體訓練效果尚有進步空間。

**7. 參考文獻**

[1] REF2013 Kaggle Data Set, [online]. Available: https://www.kaggle.com/datasets/msambare/fer2013.

[2] REFplus Kaggle Data Set, [online]. Available: https://www.kaggle.com/datasets/subhaditya/fer2013plus.

[3] RAF-DB Kaggle Data Set, [online]. Available: https://www.kaggle.com/datasets/shuvoalok/raf-db-dataset.

[4] Student Engagement Kaggle Data Set, [online]. Available: https://www.kaggle.com/datasets/joyee19/studentengagement.