

國立臺灣大學電機資訊學院資訊工程學系

碩士論文

Department of Computer Science and Information Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master Thesis

針對在多樣光线下進行車內人臉辨識設計基於三元組  
概念之訓練架構

A Triplet-based Training Architecture for In-vehicle Face  
Detection under Various Lighting Situations

莊宜蓁

Yi-Chen Chuang

指導教授：莊永裕博士

Advisor: Yung-Yu Chuang, Ph.D.

中華民國 110 年 1 月

January, 2021



國立臺灣大學碩士學位論文  
口試委員會審定書

針對在多樣光线下進行車內人臉辨識設計基於三  
元組概念之訓練架構

A Triplet-based Training Architecture for In-vehicle  
Face Detection under Various Lighting Situations

本論文係莊宜蓁君 (R07922066) 在國立臺灣大學資訊工程學  
系完成之碩士學位論文，於民國 110 年 1 月 21 日承下列考試委員  
審查通過及口試及格，特此證明

口試委員：

---

---

---

---

---

---

---

---

所 長：

---



# 誌謝

感謝...



# **Acknowledgements**

I'm glad to thank...



# 摘要

近年來隨著智慧型汽車相關技術的成熟，車內人臉辨識逐漸受到重視。然而一般的人臉偵測模型尚無法在車內人臉辨識上獲得較好的結果。其原因我們推測為車內人臉辨識會碰到較多在不同光照下的例子，而一般的人臉偵測模型著重於大多數情境下的表現結果，經常忽略在較極端光照下的例子所致。在本論文中我們試著消除不同光照對圖片造成影響。我們提出一個訓練架構來訓練出能夠對輸入資料在偵測前進行正規化處理的正規器，並將此正規器和人臉偵測器接在一起進行端對端訓練優化。我們的方法在基線做得最差的測試情境中相比基線結果進步了 47.27%。

**關鍵字：** 人臉偵測, 車內人臉偵測, 影像增強, 正規化



# Abstract

As intelligent vehicle technologies become mature, in-car face detection gradually draws everyone's attention. However, general face detection models have yet to perform good when it comes to in-car face detection. We guess the reason is that while in-car face detection has to deal with more cases from various lighting situations, general face detection tends to focus on performing good on major cases and often ignores cases under extreme lighting situations. In this thesis, we tried to remove the effect lighting had on images. We proposed a training architecture to train a normalizer to normalize input images before getting detected, jointed it with a face detector and did end-to-end training for optimization. Our method outperformed baseline by 47.27% in the test scenario that baseline performed worst on.

**Keywords:** Face Detection, In-Car Face Detection, Image Enhancement, Normalization



# Contents

口試委員會審定書	iii
誌謝	v
Acknowledgements	vii
摘要	ix
Abstract	xi
1 緒論	1
2 相關研究	3
2.1 人臉偵測 . . . . .	3
2.2 低光照下的人臉偵測 . . . . .	3
2.3 影像增強 . . . . .	4
3 問題定義與方法	5
3.1 目標與啟發 . . . . .	5
3.2 方法 . . . . .	6
3.2.1 預訓練 . . . . .	7
3.2.2 主要訓練 . . . . .	12
4 實驗設定與結果	15
4.1 資料集 . . . . .	15
4.2 實驗設定 . . . . .	16

4.3 實驗結果 . . . . .	17
4.4 消融實驗 . . . . .	18
4.4.1 在訓練時使用模擬曝光不足 / 曝光過度的資料 . . . . .	18
4.4.2 在偵測前進行正規化處理 . . . . .	19
4.4.3 使用三圖一組架構進行預訓練 . . . . .	20
4.4.4 在預訓練後對正規器進行端對端訓練優化 . . . . .	21
<b>5 總結與未來目標</b>	<b>23</b>
<b>Bibliography</b>	<b>25</b>

# List of Figures

3.1	常見人臉資料集中的圖片範例	6
3.2	車內影像範例	6
3.3	測試時從輸入圖片到獲得偵測結果的流程	7
3.4	完整的訓練流程	7
3.5	預訓練時我們嘗試過的架構	8
3.6	使用舊訓練架構進行監督式學習後之結果	8
3.7	不同 $\alpha$ 下的輸出結果之比較	10
3.8	不同光照模擬下的同一場景	11
4.1	Wider Face 資料集中的圖片範例	16
4.2	測試資料集中的圖片範例	16
4.3	我們的方法與基線之視覺結果比較	22



# List of Tables

4.1 對不同資料集的命名 . . . . .	17
4.2 基線和我們的方法之結果比較 . . . . .	17
4.3 在訓練時使用模擬曝光不足 / 曝光過度的資料對測試結果的影響 . .	18
4.4 針對在偵測前進行正規化處理與否之比較 . . . . .	19
4.5 針對使用三圖一組架構進行預訓練的效果之比較 . . . . .	20
4.6 不同設定間之結果比較 . . . . .	21



# Chapter 1

## 緒論

人臉偵測一直以來都是深度學習和電腦視覺領域中一個重要的議題。經過了幾年的研究，人臉偵測也漸趨成熟，得以被應用在更多領域中。而其中一個新興的領域便是智慧型汽車。汽車駕駛和我們的生活息息相關，一直以來各大車廠以及研究人員都在不斷研究如何讓駕駛汽車變得更加方便和安全，因而研發出各項技術，如智慧駕駛輔助、自動駕駛等等。隨著這些技術的成熟，車內影像的應用也逐漸受到關注，而對車內人臉的偵測便是一項重要的核心技術。在車內進行人臉偵測能夠協助定位出駕駛與乘客的臉部位置，並能夠接著進行後續的其他應用，如進行身分辨識、偵測駕駛臉部朝向等等，能夠使駕駛汽車更加安全。

然而在現階段，人臉偵測的研究尚無法直接應用於智慧型汽車這個領域。在測試下我們發現，一般的人臉偵測模型做在車內人臉上的表現不如我們預期。車內人臉偵測的結果對後續的相關應用十分重要，因此我們認為如何改善偵測的結果是相當值得研究的。在經過觀察之後我們發現，車內人臉偵測和一般人臉偵測相比有更多在多樣光照下的例子。一般人臉偵測的研究大多著重於提升大多數情境下的表現結果，經常忽略在較極端光照下的例子，因此其在車內人臉的偵測上便表現得較不如預期。

在本研究中，我們在進行人臉偵測前用正規器先對輸入圖片進行正規化處理(Normalization)，並提出了針對此正規器的訓練架構。我們希望能透過消除光照對輸入圖片所造成的差異，來改善模型在人臉偵測上的表現結果。

在接下來的章節中我們會陸續提到以下內容：在第 2 章，我們會介紹和本研究相關的其他研究，包含了人臉偵測、低光照下的人臉偵測和影像增強。在第 3 章，

我們會先闡述我們確定本研究中目標的過程，然後詳細說明在本研究中所使用的方法細節。在第 4 章，我們會說明我們在訓練和測試時使用的資料集和實驗設定，展示視覺和數據上的結果，並以實驗說明我們的方法中各個細節對實驗結果造成的影响。而在第 5 章，我們會對本研究做出結論並提出未來努力的目標。

# Chapter 2

## 相關研究

### 2.1 人臉偵測

人臉偵測在電腦視覺領域是個發展成熟的議題，其因能被應用在諸多領域上而有很大的重要性。從以前到現在，人們使用各種不同的方式來解決這個議題。由 P. Viola 和 M. J. Jones[16] 所提出的 Viola-Jones 目標檢測框架結合了積分圖 (Integral Image)、哈爾特徵 (Haar Feature)、自適應增強 (AdaBoost) 學習、將數個弱分類器級聯 (Cascade) 等概念，率先做到了實時性 (Real-Time) 高精度人臉偵測。而在卷積神經網路興起後，大家對人臉偵測的研究又更加熱烈。由 S. Zhang [20] 所提出的 FaceBoxes 藉由在卷積層使用較大的步伐 (Stride) 快速將輸入縮小，在盡量不影響結果下減少輸出所需的頻道數，並搭配 Faster R-CNN [13] 中核心的 RPN (Region Proposal Network) 網路和錨點 (Anchor) 的機制來做到高精度實時性的人臉偵測；由 J. Li [10] 所提出的 DSFD 強化圖片中被擷取的特徵並利用這兩組特徵來算出比單一組特徵更準確的臉部位置；由 V. Bazarevsky [1] 所提出的 BlazeFace 則使用了輕量化的網路和需要的運算處理較低的架構等，使人臉偵測能夠進一步被用於行動裝置的相關應用上。

### 2.2 低光线下的人臉偵測

人臉偵測並非只被應用於正常光照下的情境，有時候我們也會需要處理夜間等較有挑戰性的情境，而大家對於這樣的情境也各有不同的解決方案。S. W. Cho [5]

使用低光照下的圖片作為訓練資料，試圖使訓練時的情境接近測試時的情境以提升人臉偵測的準確率；M. K. Bhowmik [2] 和 J. Kang [8] 分別使用了熱紅外光攝影機和近紅外光攝影機獲取非可見光照射下的資料，以避開低光照對圖片造成影響；也有一些研究 [5, 11, 19] 選擇對圖片進行增強，試圖還原在低光照下損失的色彩與細節以提升準確率。

## 2.3 影像增強

影像增強也是一個非常熱門的議題，人們透過增強圖片來還原圖片在低光照下損失的色彩與細節，讓圖片能被進行其他後續處理。較為傳統並廣為人知的作法有直方圖均衡化 (Histogram Equalization) [6] 和伽瑪校正 (Gamma Correction) [6] 等，這些方法主要透過拉高像素間的對比來達到強調細節的效果。過去研究人員模仿人體視覺系統發展出視網膜增強算法 [9]，近期也有研究受到這個算法的啟發。由 L. Shen [15] 提出的 MSR-net 認為視網膜增強算法的架構和卷積神經網路相似，並將其演算法轉換成卷積神經網路；C. Wei [17] 的研究將圖片分為光照 (Illumination) 和反射 (Reflectance) 兩部分，把光照部分增強後再和反射部分相乘得到最終結果。除此之外，還有其他基於深度學習對影像增強的研究。J. Cai [3] 的研究將圖片分為低頻和高頻部分分別進行增強，融合兩個結果後又進行第二次增強來獲得最終結果；由 K. G. Lore [12] 所提出的 LL-net 則使用自編碼器 (Autoencoder) 對低光照圖片進行去噪和增強。

# Chapter 3

## 問題定義與方法

### 3.1 目標與啟發

本研究最終目標是希望能改善對車內人臉的偵測結果，因此首先我們先觀察為何一般人臉偵測做在車內人臉上無法獲得我們所期望的結果。透過觀察幾個較常見的人臉資料集，如 Wider Face [18] 和 AFW [21] 等，我們發現資料集中大部分都是在正常光照下的圖片 (如圖 3.1)。而與之相比，車內影像受到車輛移動的影響，會暴露於各種不同的光线下 (如圖 3.2)，使測試資料的光照更加多樣化而難以偵測出人臉。

為了解決這項問題，我們希望運用一些方法來消除光照對圖片造成的影响。首先我們注意到某個著名的傳統演算法：直方圖均衡化。在實作上它會重新分配圖中每個像素的數值使最後輸出結果的亮度在直方圖上分布均勻，有點類似將圖片的亮度進行正規化，很接近我們的需求。但這個演算法的缺點之一便是無法靈活處理每一種輸入圖片，如果碰到了有很大光線差的圖片，這個演算法並沒有辦法處理得很好。此外，使用這個演算法的話，我們便無法在事後對這部分的模型進行端對端優化。於是我們轉而考慮增強圖片。過去有不少電腦視覺的相關研究 [5, 7, 4] 在碰到光照問題時選擇對圖片進行增強處理，而我們的測試也顯示增強低光照的圖片確實能增加模型偵測出的人臉數量，因此我們原先也試過對圖片進行增強處理來解決問題。但我們發現現行的增強方法大多專注於對低光照圖片的資訊恢復，做在高光照圖片上效果並不突出。而且許多增強方法皆強調圖中的對比，而這並不一定符合我們的需求。最後我們還是回到了思考的原點，我們決

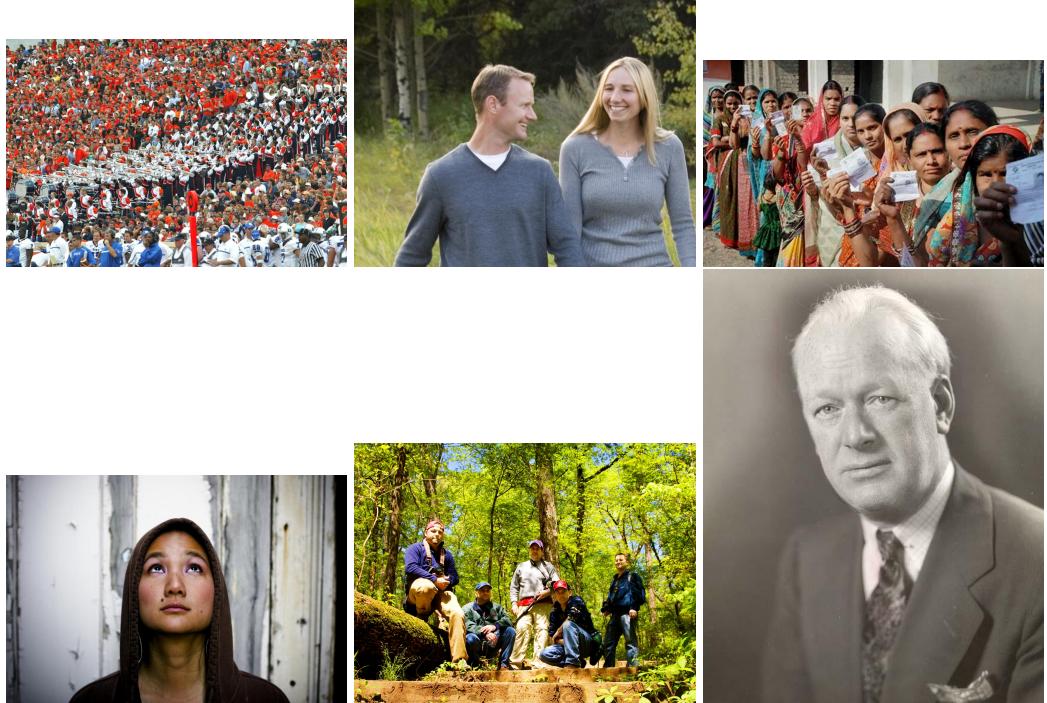


Figure 3.1: Wider Face 和 AFW 等常見人臉資料集中大部分是在正常光照下的圖片



Figure 3.2: 車內影像包含各種光照下的圖片

定試著訓練出一個模型，使其能對圖片的亮度進行正規化。

## 3.2 方法

我們的方法完整的測試流程如圖 3.3 所示。圖片在輸入模型之後會先經過正規器進行圖片的正規化處理，然後才會經過人臉偵測器算出偵測框 (Bounding Boxes) 結果。

我們完整的訓練流程則如圖 3.4 所示，分為兩個階段：預訓練和主要訓練。首先我們會先對正規器模型進行預訓練獲得初始權重，然後再將正規器和人臉偵測器接在一起進行主要訓練。以下我們分別就兩個訓練階段作詳細的說明。

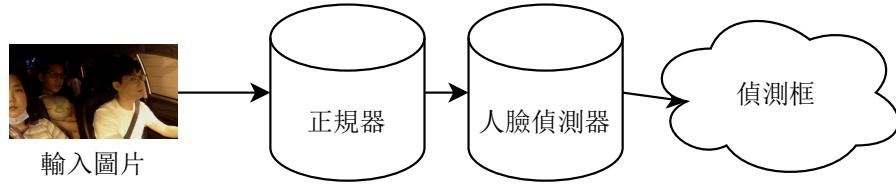


Figure 3.3: 輸入圖片會先經過正規器 (Normalizer) 處理再經過人臉偵測器算出偵測框

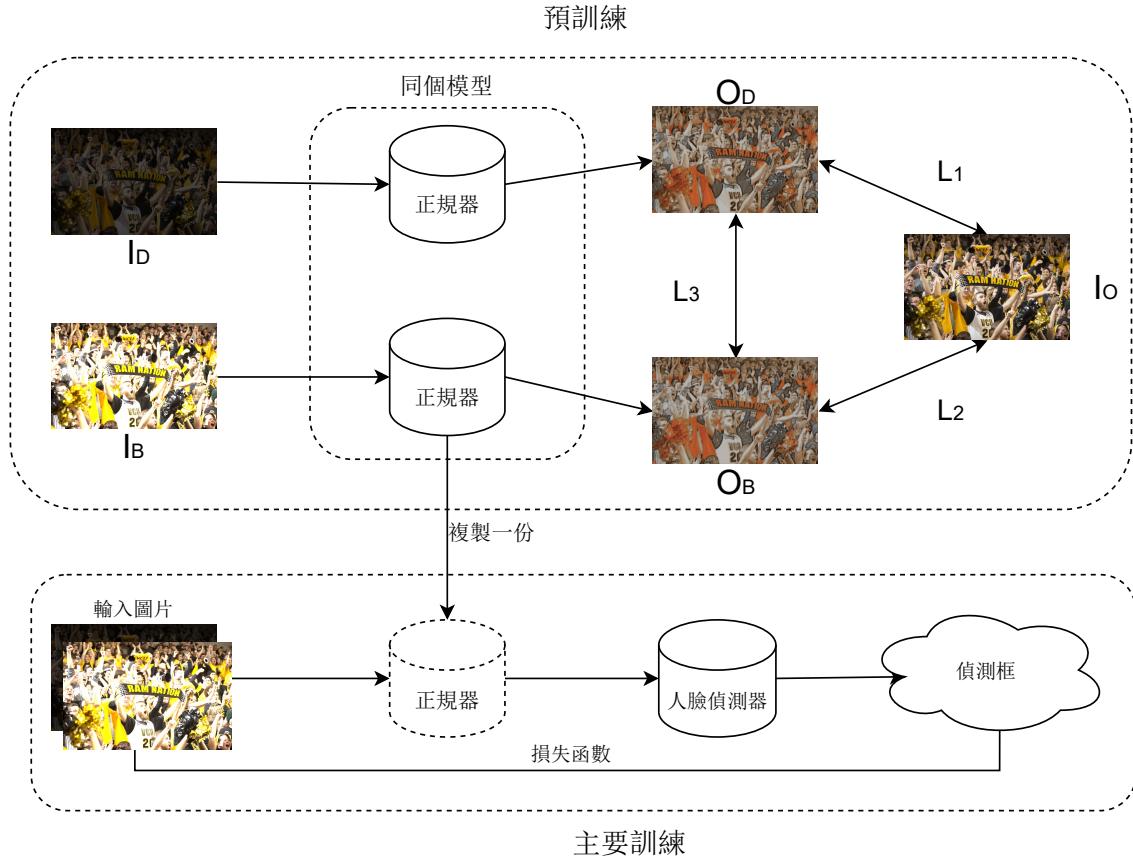


Figure 3.4: 我們的訓練流程包含了預訓練和主要訓練

### 3.2.1 預訓練

在預訓練中，我們的目的是給正規器一個好的初始權重，以便在後續的主要訓練進行優化。

對於正規器的訓練，我們原先的想法是輸入一張低 / 高光照的圖片和一張作為基準真相 (Ground Truth) 同一場景下正常光照的圖片，讓模型進行監督式學習 (如圖 3.5)，試圖把輸入圖片的光照調整至正常值。但由於在主要訓練階段我們會將正規器和人臉偵測器接在一起做端對端訓練，為了不要讓整體模型的深度太深，我們會選擇架構較簡單、參數較少的模型作為正規器的模型架構。而這樣的

選擇帶來的一個問題便是正規器的模型較弱，直接作監督式學習的結果不佳(如圖 3.6)。為了能順利訓練正規器，我們設計了一套基於三元組概念的訓練架構。以下我們分別就這套訓練架構中的損失函數、資料和所使用的模型進行說明。

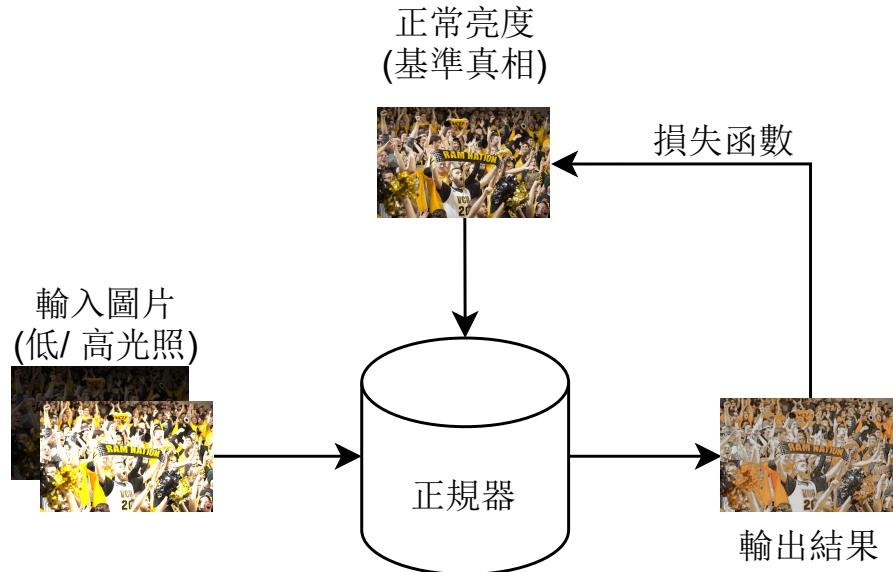


Figure 3.5: 原先我們嘗試輸入一張低 / 高光照的圖和一張作為基準真相同一場景下正常光照的圖，讓模型進行監督式學習



Figure 3.6: 我們輸入一張低 / 高光照的圖片和一張作為基準真相同一場景下正常光照的圖片，讓模型進行監督式學習，但訓練結果並不理想

## 損失函數

我們希望我們設計出的損失函數能使在不同光照下的圖片被正規化成相似的結果，以便進行後續的人臉偵測。為了達成這個目標，我們對同一場景在三個不同光照條件下的圖片進行比較。這個想法首先來自於前述的監督式學習。我們判斷以正常光照下的圖做為參考直接作監督式學習對模型而言太過困難，其理由包括了至少以下兩點：

1. 為了降低整體模型的深度，我們選擇的正規器模型架構較簡單、參數較少，是個比較弱的模型。

2. 我們用來作參考的圖雖然都是在正常光照下，但其實不同圖片間的亮度標準差不小，監督式學習難以判斷要將輸入圖片調整成什麼樣子。

為此我們決定將問題簡單化，只要求不同光照下的輸入圖片在經過處理後，圖片的亮度能收斂到某幾個值，並不要求輸出的圖片看起來應該要和同一場景在正常光照下的圖片相似。但如此一來，我們就不能直接讓模型進行監督式學習，而需要用較迂迴的方式來訓練模型。為此我們設計了一個基於三元組概念的架構，這個架構受到了 FaceNet [14] 的啟發，他們在論文中設計的損失函數會透過比較三個點之間的距離來修改參數權重。在概念上我們的損失函數定義如下：

$$L_{Total} = \alpha L_{Content} + L_{Light}$$

其中  $L_{Light}$  的用意是使不同光照下的輸入在經過處理後能有盡量一致的長相，但如果只使用  $L_{Light}$  的話，我們可能會訓練出一個不顧輸入圖片長相一律輸出同樣結果的模型。因此為了使學出來的模型能兼顧保留場景原本樣貌和消除光線影響，我們加入了  $L_{Content}$  以保留場景中的資訊。 $\alpha$  在這裡作為一個調整兩個損失函數間取捨的超參數。當  $\alpha$  越小，代表我們越傾向讓輸入圖片經過處理的結果一致，此時我們獲得的結果圖片會越來越模糊；而當  $\alpha$  越大，代表我們越強調要保留場景中的資訊，此時模型的訓練模式便會越來越接近原本的監督式學習，因乏適 (Underfitting) 而使結果圖片出現雜訊 (如圖 3.7)。而我們便必須在眾多的  $\alpha$  值中選出最適當的來進行預訓練。

我們使用一組包含三張圖片的三元組來實作上述損失函數的細節。每一組圖片包含了同一場景在曝光正常、曝光不足、曝光過度這三種情況下的圖片，以下稱  $I_O$ 、 $I_D$ 、 $I_B$ 。我們讓  $I_D$  和  $I_B$  在經過正規器後輸出的兩個結果 (以下稱  $O_D$  和  $O_B$ ) 算出兩張圖間的損失函數  $L_3$ 。為了使兩張圖之間的距離越小越好，我們在這裡使用了  $L_2$  損失函數。同時為了不使這兩個結果和原圖相差過大，我們讓  $O_D$  和  $O_B$  分別和  $I_O$  算出兩個損失函數  $L_1$  和  $L_2$ ，在這裡我們同樣使用了  $L_2$  損失函數。這



(a) 輸入圖片



(b)  $\alpha = 0.01$



(c)  $\alpha = 0.05$



(d)  $\alpha = 0.1$

Figure 3.7: 當  $\alpha$  越小，代表我們越傾向讓輸入圖片經過處理的結果一致，此時我們獲得的結果圖片會越來越模糊；而當  $\alpha$  越大，代表我們越強調要保留場景中的資訊，此時模型的訓練模式便會越來越接近原本的監督式學習，因乏適而使結果圖片出現雜訊

三個損失函數和我們在前面定義的損失函數關係如下：

$$L_{Light} = L_3 = \mathcal{L}^{\ell_2}(O_D, O_B)$$

$$L_{Content} = L_1 + L_2 = \mathcal{L}^{\ell_2}(O_D, I_O) + \mathcal{L}^{\ell_2}(O_B, I_O)$$

於是完整的損失函數定義如下：

$$L_{Total} = \alpha(L_1 + L_2) + L_3$$

## 資料

由於利用真實資料會有收集上的困難，也較難保證場景中的內容物不變，我們在此架構中所使用的三元組資料是透過演算法將原始圖片做調整，模擬出曝光不足和曝光過度這兩個情境下的圖片（示意如圖 3.8）。

模擬時進行的計算大致上分為兩類：調整圖片亮度和在圖片中增加雜訊。在模擬曝光不足的時候，我們除了調整圖片亮度之外，為了模擬夜間雜訊較多的情況會在圖片中增加雜訊。相對的，模擬曝光過度的時候，我們就只調整圖片的亮度。

當我們調整圖片的亮度時，我們會先將圖片從原本的 RGB 格式轉為 HSV 格



(a) 原始圖片

(b) 曝光不足模擬後

(c) 曝光過度模擬後

Figure 3.8: 同一場景經過不同光照模擬後的示意圖

式。這一步是為了將明度 (Value) 獨立出來以方便後續的調整。我們也曾嘗試過將圖片轉為 HSL 格式，這個格式同樣能將亮度 (Lightness) 獨立出來。但測試的時候我們發現經過 HSL 格式轉換的圖片和經過 HSV 格式轉換的圖片相比會有較大的色差，因此最後我們選擇了 HSV 格式。在經過格式轉換後，接著我們對圖片的明度進行反伽瑪校正。會有這一步是因為影像在被儲存成檔案時 (如 JPG 格式)，影像會自動套用伽瑪校正，通常套用的值是  $\gamma = 2.2$ 。因此我們要先去除伽瑪校正的影響，將影像的值調回線性。接著我們便對圖片的明度進行調整，調整過後再對明度進行伽瑪校正，最後再將圖片轉回 RGB 格式後輸出。

在圖片中增加雜訊的方式就相對單純。我們會對圖片中的每個像素增加服從於  $N(0, \sigma^2)$  的高斯雜訊後輸出。

以下分別條列兩種情境下的調整公式：

### 模擬曝光不足

1. 對每個像素  $P = (r, g, b)$  調整亮度至  $l\% (l < 1)$

$$(h, s, v) = \text{RgbToHsv}(r, g, b)$$

$$v' = ((\frac{v}{255})^{2.2} \times l\%)^{\frac{1}{2.2}} \times 255$$

$$P' = \text{HsvToRgb}(h, s, v')$$

2. 對每個調暗過後的像素  $P'$  增加高斯雜訊

$$P'' = P' + n$$

$$\text{where } n \sim N(0, \sigma^2)$$

### **模擬曝光過度**

對每個像素  $P = (r, g, b)$  調整亮度至  $l\% (l > 1)$

$$\begin{aligned}(h, s, v) &= \text{RgbToHsv}(r, g, b) \\v' &= ((\frac{v}{255})^{2.2} \times l\%)^{\frac{1}{2.2}} \times 255 \\P' &= \text{HsvToRgb}(h, s, v')\end{aligned}$$

在我們用上述的調整公式對資料進行調整後，我們就分別獲得了模擬曝光不足包含雜訊的圖片以及模擬曝光過度的圖片。

### **所使用的模型**

在此架構中我們使用了 MSR-net [15] 來作為正規器的架構。選擇這個模型的理由有以下幾項：

1. 它是一個能夠對低亮度圖片做影像增強的模型，能夠對圖片作局部亮度的調整，能夠用來對圖片進行正規化處理。
2. 我們希望這個架構能夠接受不同大小的圖片作為輸入資料，而這個模型在架構上的設計基於卷積神經網路，能夠接受不同大小的輸入資料。
3. 最後如前所述，我們在主要訓練階段會將正規器和人臉偵測器接在一起做端對端訓練，因此我們希望採用的正規器架構簡單。這個模型在尺寸上和其他影像增強的模型相比較為輕量，方便我們進行後續的訓練。

在預訓練中，我們使用基於三元組概念的架構搭配特別設計的損失函數與用演算法模擬的資料對模型進行訓練。在預訓練結束後，我們就能獲得一個能夠消除光線對圖片影響的正規器。

### **3.2.2 主要訓練**

在主要訓練中，我們將正規器和人臉偵測器接在一起做端對端訓練。我們的目標是讓來自不同光線照射下的圖片在經過正規化處理後能夠有多人臉成功被偵測出來。為了模擬不同光線照射下的情境，我們使用了上一節提到的演算法對訓練用的資料做了不同光線照射下的模擬，讓輸入資料包含曝光不足和曝光過度情境

下的圖片。

而在模型架構上，我們選擇了 FaceBoxes [20] 作為主要訓練中人臉偵測器的架構。它是一個基於卷積神經網路的人臉偵測器，透過對卷積層作優化來達到高效率高精確率人臉偵測。由於長遠來說我們希望能夠在車內進行高效率的人臉偵測，因此它對我們來說是個很好的選擇。

主要訓練結束後，我們獲得最終的人臉偵測器模型。



# Chapter 4

## 實驗設定與結果

在這一章中，我們會先介紹我們在實驗中所使用到的資料集，然後說明我們在訓練和測試時使用的設定與參數，接著展示我們進行測試所獲得在數據與視覺上的結果，最後展示對我們的方法進行消熔實驗的結果與討論。

### 4.1 資料集

我們在實驗中所使用到的資料集主要分為訓練用和測試用的資料集。

訓練時我們使用了 Wider Face [18] 作為主要的資料集。它是一個人臉資料集，其收集了各種不同情境下的人臉(如圖 4.1)，共計 32,203 張圖片、393,703 張人臉，在人臉偵測這個議題上是很經典的資料集。在我們的實驗中，我們將資料集中的圖片分別做了曝光不足和曝光過度的模擬。其中曝光不足的模擬將圖片亮度隨機調為 3%、5%、7%，並將每張圖隨機分配 1~10 的數字  $n$ ，以  $\sigma = n$  的設定對每張圖做高斯雜訊；曝光過度的模擬則將圖片亮度隨機調為 100%、250%、400%。

測試時我們則使用了自己拍攝的車內影像和人工標註的偵測框。我們在進行拍攝時將 Patriot F5 置於前擋風玻璃右上角，影像解析度為  $1280 \times 720$ ，幀率為 30 fps。影像中會出現三個人，包含駕駛、副駕駛、後座的乘客。標註時我們採用和訓練時所使用的 Wider Face 資料集相同的定義，使偵測框緊貼臉部的前額、下巴和臉頰。此資料集有以下四個情境：白天進出隧道 900 張、夜間極暗 900 張、夜間等紅燈 900 張、夜間隧道內 1000 張(如圖 4.2)。白天進出隧道是四個情境中最亮的，用來測試進出隧道時的光線劇烈變化造成的影響；夜間極暗是一個環境



Figure 4.1: Wider Face 資料集收集了各種不同情境下的人臉



Figure 4.2: 測試資料集中包含白天進出隧道、夜間極暗、夜間等紅燈和夜間隧道內四個情境

光非常微弱的情境，用來測試缺乏環境光造成影響；夜間等紅燈是一個會受到紅燈直接照射的情境，用來測試資料有色差造成影響；夜間隧道內是一個隧道光源較微弱的情境，用來測試環境光僅能照亮部分影像造成影響。在後續的實驗我們也會將這四個情境分開測試。為求說明方便，後續提到資料集時會使用表 4.1 中對資料集的命名。

## 4.2 實驗設定

訓練的流程包含預訓練和主要訓練。

在預訓練中，我們使用  $D_{Triple}$  作為輸入資料集。首先我們將每張圖片調整為  $1024 \times 1024$  的大小以配合後續訓練要求，然後將圖片切為 256 張  $64 \times 64$  的小圖

Table 4.1: 對不同資料集的命名

資料集名稱	說明	使用時機
$D_{Original}$	Wider Face 的原始資料集	-
$D_{Dark}$	將 $D_{Original}$ 做曝光不足模擬後的資料集	-
$D_{Bright}$	將 $D_{Original}$ 做曝光過度模擬後的資料集	-
$D_{Triple}$	包含 $D_{Original}$ 、 $D_{Dark}$ 、 $D_{Bright}$ 三個資料集	預訓練
$D_{Train}$	包含 $D_{Dark}$ 和 $D_{Bright}$ 兩個資料集	主要訓練
$D_{Test}$	我們拍攝的車內影像	測試

片餵進模型以  $\alpha = 0.05$  進行訓練，經過 30 萬個時期 (Epoch) 後得到正規器在主要訓練中的初始權重。接著我們進行主要訓練，在這個階段我們使用  $D_{Train}$  作為輸入資料集，把預訓練中得到的正規器和人臉偵測器接在一起做 150 個時期的端對端訓練。測試時我們使用  $D_{Test}$  作為輸入資料集，對其四個情境分別進行測試和結果評估。

### 4.3 實驗結果

以下會展示用我們的方法測試的結果和用基線測試的結果在數據和視覺上的比較。基線是用將 FaceBoxes 的架構以和原論文中同樣的資料集 ( $D_{Original}$ ) 訓練得到的模型直接對  $D_{Test}$  進行測試，並已事先測試過該模型做在論文中提到的測試資料上結果和原論文結果相似。數據上的結果比較如表 4.2，表中數據是我們使用全類平均正確率 (Mean Average Precision) 對偵測結果計算得出的數字，並已經去除部分和本研究無關的結果影響。

Table 4.2: 基線和我們的方法之結果比較

模型名稱	白天進出隧道	夜間極暗	夜間等紅燈	夜間隧道內
基線 (FaceBoxes)	63.54%	31.70%	93.35%	47.46%
我們的方法	76.25%	78.97%	97.51%	71.66%

由數據結果可以發現我們的方法得出的結果在環境光較微弱的夜間極暗、夜間隧道內情境下和基線相比分別有 47.27% 和 24.20% 的顯著進步，而在白天進出隧道、夜間等紅燈情境下則分別有 12.71% 和 4.16% 較小幅的進步。從視覺上的結果 (如圖 4.3) 也可看出我們的方法和基線相比能夠偵測出更多曝光不足的人臉，同時兼顧偵測受到正常光線照射下人臉的準確率。

## 4.4 消融實驗

在這一節，我們會探討我們方法中的每個架構與細節是否真的對車內人偵測有效。我們用  $D_{Test}$  在不同訓練設定下的模型上作實驗，來確認我們方法中的各個步驟都是有效的。首先我們會先探討在訓練時使用模擬曝光不足 / 曝光過度的資料對實驗結果的影響；接著我們會探討在輸入圖片進行偵測前先進行正規化處理對實驗結果的影響；然後我們會探討在預訓練階段使用三圖一組架構對實驗結果的影響；最後我們會探討在主要訓練階段對正規器繼續進行優化對實驗結果的影響。

### 4.4.1 在訓練時使用模擬曝光不足 / 曝光過度的資料

在我們的方法中，第一個和基線方法不同的地方在於我們訓練時不是使用  $D_{Original}$  這個收集了正常光照下圖片的資料集，而是使用了將  $D_{Original}$  模擬曝光不足情境生成的  $D_{Dark}$  和模擬曝光過度情境生成的  $D_{Bright}$  合在一起的  $D_{Train}$  作為訓練資料集。因此首先我們要先探討這件事是否有效。

由於我們的方法的架構中有正規器，以  $D_{Original}$  作為訓練資料集意義不大，同時為了去除其他部分對模型造成的影響，最後我們選擇比較**基線**和**對照組 A**這兩個不同訓練設定下的實驗結果如表 4.3。對照組 A 的模型架構只包含人臉偵測器而沒有對輸入圖片進行正規化處理，以  $D_{Train}$  作為訓練資料集。兩者差別在於主要訓練階段所使用的資料集不同。基線設定以  $D_{Original}$  作為訓練資料集；對照組 A 設定則使用了經過演算法模擬曝光不足 / 曝光過度的圖片，以  $D_{Train}$  作為訓練資料集。

Table 4.3: 在訓練時使用模擬曝光不足 / 曝光過度的資料對測試結果的影響

設定名稱	白天進出隧道	夜間極暗	夜間等紅燈	夜間隧道內
基線	63.54%	31.70%	93.35%	47.46%
對照組 A	70.97%	<b>63.16%</b>	90.75%	62.73%

從數據上的結果我們可以得知，對照組 A 設定和基線設定相比在白天進出隧道情境有 7.43% 的進步，在夜間隧道內情境有 15.27% 的進步，而在環境光非常微弱的夜間極暗情境更是有 31.46% 的進步。我們認為這代表使用較極端光線照射

下的圖片作為輸入資料有助於讓偵測器在訓練階段認識更多在不同光線照射下的人臉，使我們在測試階段能夠偵測出更多原本受到光照影響而不認識的人臉。

#### 4.4.2 在偵測前進行正規化處理

在我們的方法中，我們在偵測對輸入圖片進行了正規化處理。接下來我們會探討這一步對實驗結果的影響。我們比較**對照組 A**、**對照組 B**、**對照組 C**、**我們的方法**這四個不同訓練設定下的實驗結果如表 4.4。

對照組 A 如上一小節所述，是以  $D_{Train}$  作為訓練資料集，整個模型架構只包含人臉偵測器，沒有對輸入圖片進行正規化處理的設定。對照組 B 在預訓練階段盡可能還原了我們所使用的正規器模型架構 MSR-net [15] 的訓練設定，以  $D_{Dark}$  作為預訓練的訓練資料集，以用視網膜增強算法對  $D_{Dark}$  的圖片進行處理後的結果圖片作為基準真相，對正規器進行預訓練。接著在主要訓練階段，我們以  $D_{Train}$  作為訓練資料集，對正規器和人臉偵測器進行端對端訓練。對照組 C 在預訓練階段使用了第 3 章圖 3.5 所提到的架構，以包含了  $D_{Dark}$  和  $D_{Bright}$  的  $D_{Train}$  作為預訓練資料集、 $D_{Original}$  作為基準真相，讓正規器進行監督式學習。接著在主要訓練階段，我們以  $D_{Train}$  作為訓練資料集，對正規器和人臉偵測器進行端對端訓練。我們的方法在預訓練階段以包含了  $D_{Dark}$ 、 $D_{Bright}$ 、 $D_{Original}$  的  $D_{Triple}$  作為預訓練資料集，採用三圖一組的架構(如圖 ??)，以迂迴的方式對正規器進行預訓練。接著在主要訓練階段，我們以  $D_{Train}$  作為訓練資料集，對正規器和人臉偵測器進行端對端訓練。

除了對照組 A 之外的其他設定分別採用了不同架構或設定對正規器進行預訓練，但三者之間共通的是都有對輸入圖片進行正規化處理，僅對照組 A 沒有。

Table 4.4: 針對在偵測前進行正規化處理與否之比較

設定名稱	白天進出隧道	夜間極暗	夜間等紅燈	夜間隧道內
對照組 A	70.97%	63.16%	90.75%	62.73%
對照組 B	74.66%	81.21%	95.42%	69.54%
對照組 C	73.56%	76.02%	96.59%	68.73%
我們的方法	76.25%	78.97%	97.51%	71.66%

從數據上的結果我們可以看到，對輸入圖片進行過正規化的設定無一例外都獲得了比對照組 A 更好的結果。這能夠證明對輸入資料進行正規化處理對人臉

偵測有幫助，並且在環境光較微弱的情境下這個影響更加明顯，如在夜間極暗情境下，對照組 B、對照組 C、我們的方法和對照組 A 相比，分別有 18.05%、12.86%、15.81% 的進步。我們認為這代表正規化處理成功將不同光照下的圖片內容進行調整使其有相似的長相，讓後續的偵測過程更加順利。特別要注意的是，雖然我們的方法和對照組 B 相比在夜間極暗情境下有 2.24% 的些微退步，但這個情境的環境光極度微弱，在部分圖片中兩者都只有抓出人臉的一部分，而有些偵測框碰巧跨過了  $IOU >= 50\%$  的門檻。實際上這兩個設定在這個情境下視覺上相差並不大。

#### 4.4.3 使用三圖一組架構進行預訓練

接著我們希望能證明，我們在預訓練階段所使用的三圖一組架構對人臉偵測是有幫助的。為此我們比較**對照組 C** 和**我們的方法**這兩個不同訓練設定下的實驗結果如表 4.5。兩個設定的詳細訓練設定在上一小節已經詳述，這裡不再贅述。兩個設定都會對輸入圖片進行正規化處理，也都有對正規器進行預訓練，預訓練階段所用到的資料集也相同。它們唯一的差別在於預訓練階段所使用的訓練架構。對照組 C 直接讓正規器進行監督式學習，而我們的方法則使用三圖一組架構迂迴地訓練正規器。

Table 4.5: 針對使用三圖一組架構進行預訓練的效果之比較

設定名稱	白天進出隧道	夜間極暗	夜間等紅燈	夜間隧道內
對照組 C	73.56%	76.02%	96.59%	68.73%
我們的方法	76.25%	78.97%	97.51%	71.66%

數據上的結果顯示，即使兩個設定都有對正規器進行預訓練，在後續的主要訓練階段也都有進行端對端訓練優化，且在預訓練階段用到的資料集、主要訓練階段使用的資料集和訓練設定等都一致，我們的方法卻有比較好的結果，在四個測試情境下分別進步了 2.69%、2.95%、0.92%、2.87%。在第 3 章我們也提過，由於主要訓練階段的需要，我們在預訓練階段所使用的模型架構有較少的參數也較弱，因此對照組 C 的架構無法順利訓練出我們想要的結果。但在我們更動架構後，我們成功獲得了較佳的偵測結果。我們認為這代表預訓練時三圖一組的架構能夠有效降低模型學習如何對圖片進行正規化處理的難度，使訓練結果更容易收

斂。

#### 4.4.4 在預訓練後對正規器進行端對端訓練優化

最後我們探討在主要訓練階段對正規器進行端對端訓練優化的必要性。

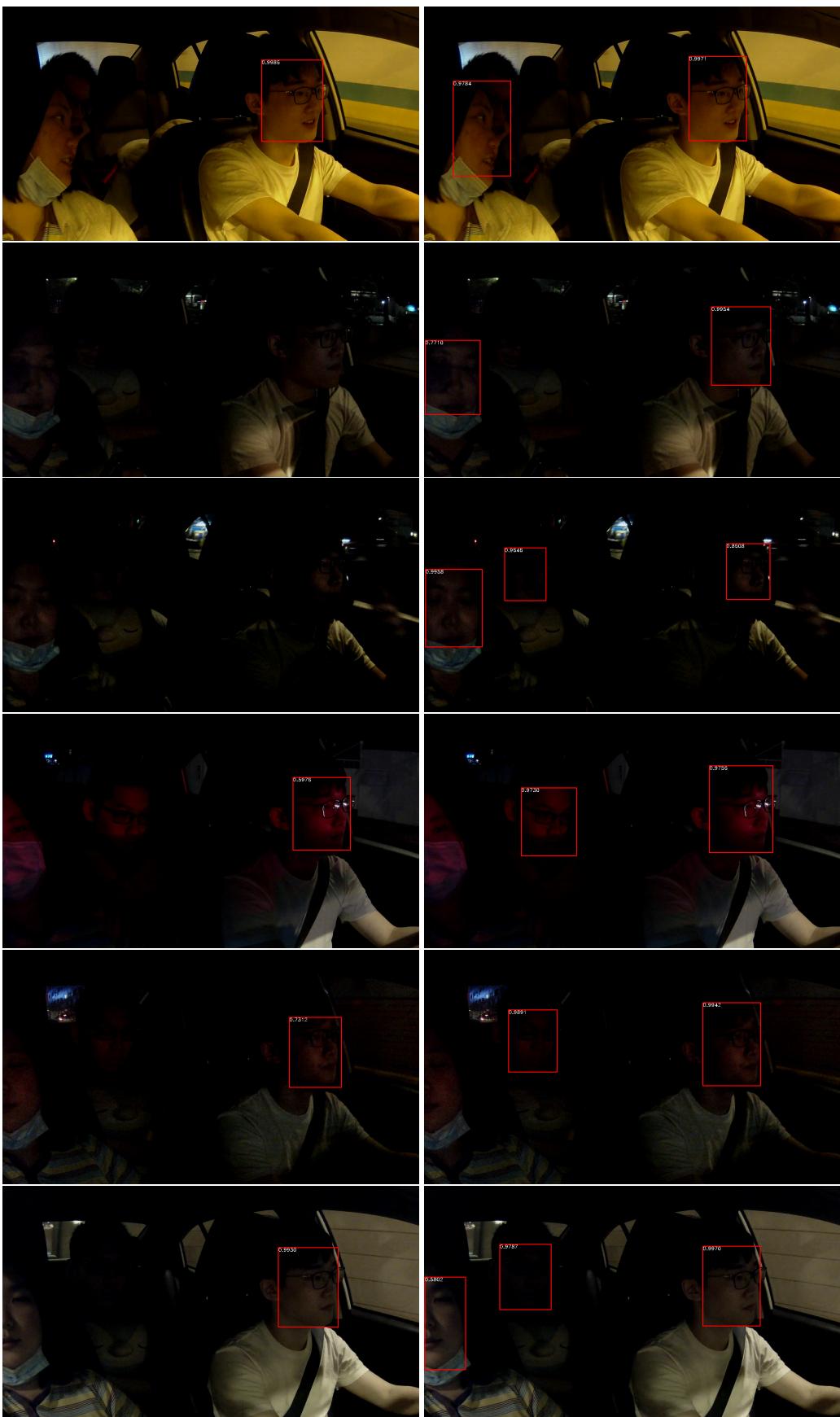
我們分別比較兩對設定：**對照組 D 和對照組 B、對照組 E 和我們的方法**以及**對照組 A**這幾個不同訓練設定下的實驗結果如表 4.6。對照組 A、對照組 B、我們的方法的詳細設定在前面都已經詳述，這裡不再贅述。對照組 D 在預訓練階段和對照組 B 完全一致，盡可能還原了 MSR-net 的訓練設定進行預訓練。但在主要訓練階段，我們凍結了正規器的權重，只更新人臉偵測器部分的權重。對照組 E 在預訓練階段和我們的方法完全一致，使用了三圖一組架構對正規器進行預訓練。但在主要訓練階段，我們凍結了正規器的權重，只更新人臉偵測器部分的權重。兩對設定有一個共通點，就是都有一個設定（對照組 D 和對照組 E）在主要訓練時凍結了正規器的權重，只訓練人臉偵測器本身。對照組 A 在這個比較中則作為沒有用到正規器的對照組。

Table 4.6: 不同設定間之結果比較

設定名稱	白天進出隧道	夜間極暗	夜間等紅燈	夜間隧道內
對照組 A (無正規器)	70.97%	63.16%	90.75%	62.73%
對照組 D (未優化)	73.15%	60.13%	88.67%	63.88%
對照組 B (有優化)	74.66%	81.21%	95.42%	69.54%
對照組 E (未優化)	22.61%	41.13%	83.31%	30.35%
我們的方法 (有優化)	76.25%	78.97%	97.51%	71.66%

數據上的結果顯示，光是進行正規器的預訓練是不夠的。如果不進行後續的優化，結果可能反而會比不進行正規化處理還要差。我們推測這是由於正規器對圖片進行的調整雖然拉近了不同光照下圖片間的距離，卻增加了一些本不應存在的雜訊或使圖片產生了不該有的色偏，使得偵測器無法順利偵測出人臉。但預訓練依然給了正規器一個好的初始權重，讓模型知道這樣調整圖片可以拉近彼此間的距離。而端對端訓練的意義便在於對正規器做的事情進行調整與優化，使它在拉近圖片間距離的同時顧及偵測器對人臉偵測的需求。

在經過以上的測試後，我們確認了我們的方法中的各個步驟都是有效的。



(a) 基線的測試結果

(b) 我們的方法的測試結果

Figure 4.3: 比較我們的方法與基線之視覺結果可發現我們的方法能偵測出更多曝光不足的人臉

# Chapter 5

## 總結與未來目標

在本研究中，我們提出了基於三元組概念的預訓練架構來訓練出能將不同光線照射下的圖片正規化成相似結果的正規器，並將正規器和人臉偵測器接在一起做端對端訓練優化來提升人臉偵測的準確率。實驗表明我們的方法對正規器的訓練有幫助，也提升了整體人臉偵測的準確率。

未來我們能夠改進的地方包含了以下幾點：我們能夠修改三元組訓練架構中的損失函數，使其能將圖片的正規化處理做得更好；我們還能修改模擬曝光不足、曝光過度圖片的演算法，使其能夠更貼近真實圖片的樣貌以利訓練進行。



# Bibliography

- [1] V. Bazarevsky, Y. Kartynnik, A. Vakunov, K. Raveendran, and M. Grundmann. Blazeface: Sub-millisecond neural face detection on mobile gpus. *arXiv preprint arXiv:1907.05047*, 2019.
- [2] M. K. Bhowmik, K. Saha, S. Majumder, G. Majumder, A. Saha, A. N. Sarma, D. Bhattacharjee, D. K. Basu, and M. Nasipuri. Thermal infrared face recognition—a biometric identification technique for robust security system. *Reviews, refinements and new ideas in face recognition*, 7, 2011.
- [3] J. Cai, S. Gu, and L. Zhang. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Transactions on Image Processing*, 27(4):2049–2062, 2018.
- [4] C. Chen, Q. Chen, J. Xu, and V. Koltun. Learning to see in the dark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3291–3300, 2018.
- [5] S. W. Cho, N. R. Baek, M. C. Kim, J. H. Koo, J. H. Kim, and K. R. Park. Face detection in nighttime images using visible-light camera sensors with two-step faster region-based convolutional neural network. *Sensors*, 18(9):2995, 2018.
- [6] R. C. Gonzales and R. E. Woods. Digital image processing, 2002.
- [7] P. Kalaiselvi and S. Nithya. Face recognition system under varying lighting conditions. *IOSR Journal of Computer Engineering (IOSRJCE)*, 14(3):79–88, 2013.

- [8] J. Kang, D. V. Anderson, and M. H. Hayes. Face recognition in vehicles with near infrared frame differencing. In *2015 IEEE Signal Processing and Signal Processing Education Workshop (SP/SPE)*, pages 358–363. IEEE, 2015.
- [9] E. H. Land. The retinex theory of color vision. *Scientific american*, 237(6):108–129, 1977.
- [10] J. Li, Y. Wang, C. Wang, Y. Tai, J. Qian, J. Yang, C. Wang, J. Li, and F. Huang. Dsfd: dual shot face detector. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5060–5069, 2019.
- [11] J. Li, D. Zhang, K. Zhang, K. Hu, and L. Yang. Real-time face detection during the night. In *2017 4th International Conference on Systems and Informatics (ICSAI)*, pages 582–586. IEEE, 2017.
- [12] K. G. Lore, A. Akintayo, and S. Sarkar. Llnet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognition*, 61:650–662, 2017.
- [13] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.
- [14] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.
- [15] L. Shen, Z. Yue, F. Feng, Q. Chen, S. Liu, and J. Ma. Msr-net: Low-light image enhancement using deep convolutional network. *arXiv preprint arXiv:1711.02488*, 2017.
- [16] P. Viola and M. J. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.
- [17] C. Wei, W. Wang, W. Yang, and J. Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018.

- [18] S. Yang, P. Luo, C.-C. Loy, and X. Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5525–5533, 2016.
- [19] W. Yang, Y. Yuan, W. Ren, J. Liu, W. J. Scheirer, Z. Wang, T. Zhang, Q. Zhong, D. Xie, S. Pu, et al. Advancing image understanding in poor visibility environments: A collective benchmark study. *IEEE Transactions on Image Processing*, 29:5737–5752, 2020.
- [20] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li. Faceboxes: A cpu real-time face detector with high accuracy. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9. IEEE, 2017.
- [21] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2879–2886. IEEE, 2012.