關於運用多教师知识蒸馏輕量化人脸關键點檢測器研究計畫

（一）研究主題：運用多教师知识蒸馏輕量化人脸關键點檢測器

（二）研究背景

人臉關鍵點檢測（Facial Landmark Detection）是人臉識別、人臉姿態估計、人臉表情分析等後續應用的關鍵步驟，目的是在人臉圖像定位出預定義的關鍵位置，如眼睛、眉毛、鼻子、嘴和下巴。深度學習的發展大幅度提升了人臉關鍵點檢測的精度，但是目前高精度的檢測算法大多采用大型骨幹網絡，如Hourglass [1]、ResNet [2] 和 HR-Net [3]等，這些模型參數大、計算成本高和處理速度慢的模型難以滿足性能較差設備的需求。因此一些模型尺寸小、計算成本低、性能高的神經網絡架構也被設計出來以實現快速且準確的需求，包括 MobileNet [4] 和 ShuffleNet [5]。 然而，將這些輕量級模型直接應用於面部對齊可能會嚴重損害性能。 因此，設計一個快速且精確的面部標誌檢測器仍然是一個具有挑戰性和意義的問題。本研究計畫運用基於多教師蒸餾網絡結合channel-wise和對抗性學習的方法輕量化HR-Net網絡，以實現高精度輕量級的人臉關鍵點檢測算法。

（三）文獻探討

1.人脸关键点检测

近年來，基於深度學習的人臉關鍵點檢測框架[6]、[7]、[8]比傳統的算法在準確性和效率方面顯示出顯著優勢。 目前這種深度學習算法可以分為兩類：坐標回歸和熱圖回歸。

坐標回歸方法使用回歸模型直接從輸入圖像中預測地標的坐標，而不依賴於外觀模型。 這些方法 [9,10－11] 通常使用從粗到精的方式迭代更新形狀。 [10] 預訓練了一個特徵提取網絡，以從全局面部特徵中學習局部面部特徵，從而提高了面部對齊精度。 DAC-CSR [11] 將人臉分成多個域來訓練特定域的級聯形狀回歸（CSR）。 一些回歸方法 [12-13] 也學習基於形狀索引特徵的回歸模型，這些特徵最初是在 ESR [12] 中提出的。 它使用平均形狀作為初始形狀，並通過基於初始形狀周圍提取的局部特徵預測偏移來逐漸更新地標。 吳等人 [13] 認為不同的臉型應該有不同的回歸函數。 因此，他們提出的模型可以根據當前人臉形狀自動改變回歸參數，以更好地逼近真實形狀。

熱圖回歸方法 [8] 通過在通道上生成高斯分布來獲得熱圖； 預測熱圖上響應最高的點可能是預測。Wang等人[14] 設計了熱圖回歸的損失函數，對前景像素的懲罰更大，對背景像素的懲罰更小。Chandran等人則 [15] 提出了一種基於注意力的空間金字塔網絡，該網絡可以提取不同尺度和分辨率的特徵，從而在不犧牲圖像分辨率和質量的情況下，結合了面部特徵檢測的全局和局部特徵信息，進一步提高了關鍵點檢測的準確性。 Wang等人[3] 在2020年提出了深度高分辨率表徵學習（HRNet），HRNet採用了多分支的並行結構，使得網絡可以同時保留多個分辨率的特徵信息，從而提高了網絡對細節的感知能力。

熱圖回歸方法可以取得更好的性能，但需要較深的網絡和較多的參數，導致計算複雜，在配備有限的內存和計算設備（例如手機和機器人）上處理速度較慢。

2.知识蒸馏

知識蒸餾是一種流行的模型壓縮技術，它涉及將知識從複雜的大型模型（教師網絡）轉移到更簡單的小型模型（學生網絡），同時保持大型模型的性能。 多年來，已經提出了幾種知識蒸餾的變體，每一種都有其獨特的優點和局限性。

Hinton等人[16]在2015年提出了知識蒸餾的概念，並將其應用於圖像分類任務中。他們通過將大型深度神經網絡的知識傳遞給小型網絡，成功地實現了對小型網絡的精度提升。知識蒸餾背後的關鍵思想是，經過訓練的「教師」網絡輸出的軟概率包含更多關於數據點的信息，而不僅僅是類標籤。學生網絡模仿這些概率能讓學生網絡吸收一些教師發現的知識，而不僅僅是訓練標籤中的信息。在隨後的幾年中，知識蒸餾產生了更多的變體。

Wang等人[17]在2018 提出了一種新的知識蒸餾方法，即將生成對抗網絡（GAN）應用於知識蒸餾中。具體地，該方法使用一個生成器網絡和一個鑒別器網絡來學習教師網絡的知識，並將學到的知識傳遞給一個更輕量級的學生網絡。在該方法中，生成器網絡通過學習從學生網絡的隨機噪聲到教師網絡特徵的映射來生成特徵，而鑒別器網絡則負責區分教師網絡的特徵和生成器網絡生成的特徵。

Finogeev等人[18] 則提出运用基於GAN方法的知識蒸餾來提高目標檢測的速度和精度。該方法利用教師網絡和學生網絡之間的對抗損失，通過對真實圖像和生成圖像之間的差異進行優化，來傳遞教師網絡的知識給學生網絡。

2021 年，Zhou等人[19]則提出了一種稱為「channel-wise knowledge distillation」的技術，將知識從教師模型的中間特徵圖轉移到學生模型。 這種方法涉及計算特徵圖的均值和方差並將它們傳輸到學生模型。它擁有更高的模型壓縮率，傳統的知識蒸餾方法只關注如何蒸餾模型的權重參數，而「channel-wise knowledge distillation」方法還關注了通道維度上的信息，可以進一步壓縮模型，減少模型的參數量和計算量。同時，它可以幫助學生模型更好地學習到教師模型在通道維度上的特徵表示，從而提升模型的精度。

S You 等人[20]在 2017 年發表的論文「Learning from multiple teacher networks」中提出了一種多教師知識蒸餾的方法，利用多個教師網絡來提高學生網絡的性能。 作者認為使用多個教師幫助網絡可以解決使用單個教師網絡中的過度擬合或偏差等局限性。作者引入了一種損失函數，鼓勵學生網絡同時匹配所有教師網絡的預測，並表明該方法可以顯著提高學生網絡在各種分類任務上的性能。該論文還提出了一種擴展方法，允許根據每個教師網絡的表現為其分配不同的權重，並表明這可以進一步提高學生網絡的準確性。

總體而言，知識蒸餾已成為模型壓縮的強大工具，其各種擴展已證明其在各種應用中的有效性。本研究計劃運用知識蒸餾及其變體的思想，更好地將大模型的知識遷移到更高效的小模型中。

3.對抗性學習

對抗性學習是近年來機器學習領域備受關注的研究方向之一。它的主要目標是通過對抗樣本來增強機器學習算法的魯棒性，以便能夠在面對未知數據時保持良好的表現。隨著對抗性攻擊技術的不斷發展，對抗性學習也在不斷地演進。

早期的對抗性學習主要集中在解決分類問題中的對抗性攻擊。Szegedy等人[21]在2013年提出了一種稱為「對抗性樣本生成」（Adversarial Examples）的方法，該方法可以生成一些微小的擾動，將原始數據樣本誤導為錯誤的分類結果。此後，Goodfellow等人[22] 在2015年提出了一種通用的對抗性攻擊方法，即生成對抗網絡（GAN）。GAN模型利用對抗的方式來訓練兩個神經網絡：生成器和判別

器。生成器負責生成對抗樣本，而判別器則試圖區分對抗樣本和真實樣本。這種方法不僅可以用於分類問題，還可以應用於圖像生成和語音合成等其他領域。

除了對抗樣本生成和對抗性攻擊，對抗性學習還涉及到了對抗性訓練、對抗性防禦和對抗性評估等方面的研究。

Kurakin等人[23]提出了一種新的對抗性訓練方法PGD，即在模型訓練過程中，使用一種基於投影梯度下降的算法來生成對抗性樣本，並將這些對抗性樣本與原始樣本一起用於模型訓練。通過不斷迭代，PGD可以使模型在對抗性攻擊下更具魯棒性。

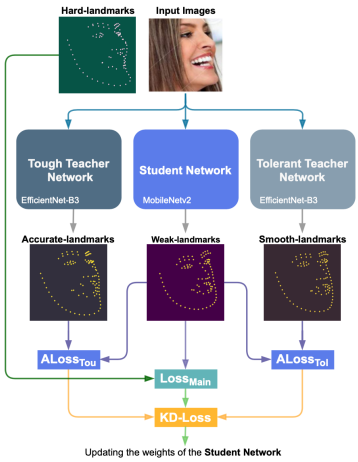
除此之外，許多工作也注重將對抗性學習應用於其他計算機視覺任務，包括人臉檢測、圖像分割和人體姿勢估計。YANG等人[24] 在2018年設計了一種新穎的多源鑑別器來區分預測的 3D 姿勢與地面實況，這有助於強制姿勢估計器生成符合人體測量學的有效姿勢，该论文还設計了一個幾何描述符用于計算成對的相對位置和身體關節之間的距離，作為鑑別器的新信息源。

4.人臉關鍵點檢測輕量化方法

儘管知識蒸餾已被廣泛用於壓縮 CNN 分類器，但少有文獻研究如何將這種技術應用於檢測任務，而基於知識蒸餾的人臉關鍵點檢測輕量化方法的相關研究更是少有。在分類器的情況下，模型的輸出是單個標籤，教師模型可以指導學生模型學習以高精度預測相同的標籤，因此知識蒸餾效果很好。

然而，對於檢測任務而言，檢測器的輸出通常包含一些來自前景類的樣本，並且主要由來自背景類的樣本組成。因此，簡單地模仿最後一層的所有輸出（如在分類和度量學習任務中）將導致模型性能不佳。且模型的輸出不是單個標籤，而是一組邊界框及其對應的標籤。 這使得知識蒸餾更具挑戰性，因為很難以確保輸出邊界框準確和精確的方式將知識從教師模型轉移到學生模型。

對於這個問題，G Chen[25]提出了加權交叉熵損失來解決將 KD 應用於基於兩階段的檢測器時的類不平衡問題。.



圖表 1Pourramezan Far等人的基於知识蒸馏流程圖

而對於將知識蒸餾運用在人脸关键点检测的例子，Pourramezan Far等人[26]在2022年提出了的基於多教师知识蒸馏人脸关键点检测模型，该文献提出了一種新的損失函數來訓練用於面部標誌檢測的輕量級學生網絡，并定義了 Mean-landmark、Soft-landmarks 和 Hard-landmarks 術語，並使用它們來訓練多教师知识蒸馏网络。然而，作者同時也指出新的學生網絡雖然相比MobileNetv2的性能有所提高，但是精度方面較先進的模型還有一定差距。

本研究未來慾借鑑前輩的思路，運用新的損失函數和更先進的模型蒸餾出更高精度輕量化的人臉檢測模型。

（四）研究目的

因此，本研究未來方向旨在應用知識蒸餾技術輕量化人臉關鍵點檢測模型，以實現更高效且精准的人臉關鍵點檢測器。具體而言，本研究目的包括：

1.探索一種基於Multi-Teacher Distillation (MTD)網絡的蒸餾方法，結合channel-wise Distillation 和對抗性學習技術，用於輕量化HR-Net網絡以實現高精度輕量化的人臉關鍵點檢測算法。

2. 通過對不同粒度特徵的知識提取和傳遞，以及對channel-wise信息的關注和整合，提高人臉關鍵點檢測模型的精度和魯棒性，並在減少模型參數和計算複雜度的同時，提高模型的運行效率和推理速度。
3. 探索對抗性學習在臉關鍵點檢測模型上的應用，以進一步提高模型的泛化能力和抗干擾能力。
4. 評估所提出方法的效能，並與現有的其他輕量級人臉關鍵點檢測模型和本研究的人臉關鍵點檢測模型進行比較。

（五）研究方法

1. 骨幹架構（Backbone Architecture）

設計合適的骨幹架構對於輕量化人臉關鍵點檢測很重要。目前，大多數精度良好的基於 CNN 的方法都採用較為複雜的主幹。為了捕獲更多信息和改進特徵表示，這些繁瑣的網絡通常包含大量重復的塊和通道，這可能會引入冗余的模型參數和昂貴的計算成本。

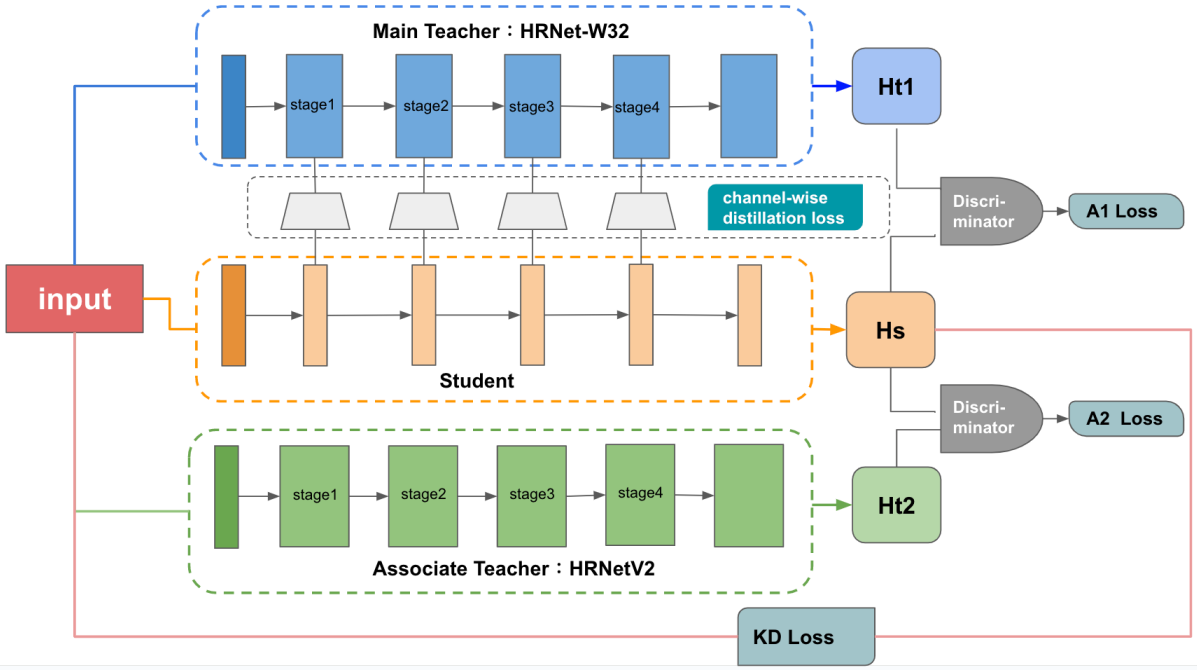


圖2 本研究的整體框架圖

教師網絡：

根據[27]的整體統計，HR-Net [3]在许多人臉關鍵點检测的基准数据集上取得了最先进的结果。 但同時它也非常消耗CPU和GPU的性能，很难直接应用于智能手机等资源受限的设备。 受 MobileNet [24] 的启发，我认为通过將適當的模型輕量化改進結合知识蒸馏，可以获得模型尺寸小、推理速度快且盡量保持精度的人臉關鍵點檢測器。

在HRNet的類別中， HRNet-W48 具有最大的寬度（即通道數）和深度，但需要更多的計算資源。HRNet-W18 是最輕量級的 HR-Net 模型，但相應的精度損失也更大。HRNet-W32 是一個折中的選擇，提供了適度的準確性和計算資源的平衡。由此，我採用 HRNet-W32 作為我教師網絡的基礎。

學生架構：

平衡骨幹架構的深度和寬度是優化性能的常用方法。 減少通道數是一種簡單而有效的修改，所以本研究計畫通過實驗，調整HRNet的寬度和深度並尋找到最能平衡模型參數和預測性能的模型參數，以簡化HRNet模型作為學生網絡。

2.知識蒸餾架構

2.1多教師知識蒸餾

基於不同結構的教師網絡有助於學生網絡從 ground truth labels 中捕獲更多互相補充的圖像特徵，並從不同的角度提高對圖像辨識性能，增加學生模型的魯棒性。

本研究計劃提出由兩個教師網絡和一個學生網絡構成的多教師知識蒸餾，具體而言，將HRNet的不同變體分別作為兩個教師模型，分別是HRNet-W32和HRNetV2模型，這種設計可以充分利用HRNet-W32在輕量化方面的優勢和HRNetV2在多尺度特徵融合方面的優勢，從而更好地指導學生模型學習。

對於損失函數的選擇，傳統知識蒸餾運用的softmax交叉熵損失只能用於單標籤方法，難以傳遞結構信息。與文獻[25]類似，本研究計畫用MSE損失函數作為標準損失函數，用來對比學生網絡預測的熱圖與真實關鍵點的熱圖之間的偏差。

2.2 channel-wise distillation

在基礎的知識蒸餾中，單純的輸出層的損失對比未能考慮教師網絡的更多中間細節，所以本研究計劃進一步在HRNet-w32和學生網絡之間加入channel-wise distillation [19]。具體而言，channel-wise distillation利用空間注意力將教師模型的注意力信息轉移到中間層的學生模型。我將每個通道的特徵圖的注意力信息視為知識，學生和老師會分別從各自的特徵圖中計算出每個通道的注意力信息，然後老師會監督學生學習每個通道的注意力信息，並將注意力信息傳遞給學生。

老師和學生的通道數通常是不匹配，我計劃使用卷積層來將學生的特徵圖維度先提升到和老師一樣的通道數，然後進行Channel-wise Distillation [19] 並計算出CD loss。

2.3對抗性學習

在輸出層將教師網絡的知識遷移到學生網絡的方法選擇上，MSE損失函數是存在一些缺點，首先，它僅關注了預測結果與真實結果之間的數值差異，而沒有考慮到它們在空間上的關係，因此可能無法準確地保留教師網絡的空間結構信息。其次，MSE損失函數對於教師網絡的錯誤可能會敏感，從而導致學生網絡產生錯誤。

本研究計畫提出對抗性學習來傳遞教師網絡的知識，以期望解決MSE損失函數的一些缺點。我計劃用CNN 分類器作為鑒別器，用於區分熱圖是從教師模型還是學生模型生成的，並幫助學生網絡進行對抗性訓練。這樣能夠更好地模擬數據分布，提高模型對於對抗樣本的魯棒性。

2.4全損失函數

為了將不同尺度的結構信息從教師網絡傳輸到學生網絡，模型運用channel-wise distillation將注意力信息傳遞給學生，並結合多教師從不同角度遷移知識，以及運用對抗性學習，從而使學生網絡學習到更豐富的面部細節。學生網絡訓練優化的總損失函數可表示為：

$$L_{total} = \lambda_{KD} L_{KD} + \lambda_{AKD}(L_{A1} + L_{A2}) + \lambda_{CD} L_{CD}$$

整个目标函数由真實人臉關鍵點生成的熱圖H與學生網絡預測的熱圖H_s之間的MSE損失L_{KD}、channel-wise 知識蒸餾損失L_{CD}和對抗性學習的損失$L_{A1} + L_{A2}$組成。其中λ_{KD}、λ_{AKD}和λ_{CD}是用於重新加權三個損失項的平衡參數。

3.訓練策略

為驗證本算法的有效性，本研究將會在主流WFLW[28]、300W[29]數據集上進行實驗。WFLW數據集是在非受限條件下採集的數據集，圖像中存在較大的姿態變化、誇張的表情和嚴重的遮擋，並標註有98個人臉關鍵點。300W數據集則由HELEN、LFPW、AFW和IBUG數據集組成，每張人臉圖像有68個標註的人臉關鍵點，泛應用於人臉關鍵點檢測中。其中除了IBUG數據集，其他數據集的圖像在野外環境中採集，可能存在姿態變化、表情變化和部分遮擋的情況。

為更好地訓練學生網絡，本研究計畫首先使用數據集對教師網絡進行訓練，得到高精度的教師網絡，而後固定教師網絡的參數，利用MSE損失函數、channel-wise 知識蒸餾損失函數和對抗性學習，在教師網絡的指導下，對學生網絡進行訓練，使總損失最小。

（六）研究預期結果

本研究旨在探索一種基於Multi-Teacher Distillation (MTD)網絡的蒸餾方法，結合channel-wise Distillation 和對抗性學習技術，用於輕量化HR-Net網絡以實現高精度輕量化的人臉關鍵點檢測算法。預期結果包括：

1）成功實現基於MTD網絡的蒸餾方法，通過對不同粒度特徵的知識提取和傳遞，以及對channel-wise信息的關注和整合，提高人臉關鍵點檢測模型的精度和魯棒性。

2）通過對抗性學習技術的應用，進一步提高模型的泛化能力和抗干擾能力。

3）通過減少模型參數和計算複雜度，提高模型的運行效率和推理速度。

4）通過對所提出方法的效能進行評估，與現有的其他輕量級人臉關鍵點檢測模型和本研究的人臉關鍵點檢測模型進行比較，證明所提出的方法的優越性。

因此，本研究的預期結果是在實現高精度的輕量化人臉關鍵點檢測算法的同時，提高模型的效率和魯棒性，為人臉關鍵點檢測任務提供更好的解決方案。

（七）參考文獻

[1] NEWELL, Alejandro; YANG, Kaiyu; DENG, Jia. Stacked hourglass networks for human pose estimation. In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*. Springer International Publishing, 2016. p. 483-499.

[2] HE, Kaiming, et al. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016. p. 770-778.

[3] WANG, Jingdong, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020, 43.10: 3349-3364.

[4] SANDLER, Mark, et al. Mobilenetv2: Inverted residuals and linear bottlenecks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018. p. 4510-4520.

[5] ZHANG, Xiangyu, et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018. p. 6848-6856.

[6] DONG, Xuanyi, et al. Style aggregated network for facial landmark detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018. p. 379-388.

[7] FENG, Zhen-Hua, et al. Wing loss for robust facial landmark localisation with convolutional neural networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018. p. 2235-2245.

[8] WU, Wayne, et al. Look at boundary: A boundary-aware face alignment algorithm. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018. p. 2129-2138.

[9] TRIGEORGIS, George, et al. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016. p. 4177-4187.

[10] PARK, Byung-Hwa; OH, Se-Young; KIM, Ig-Jae. Face alignment using a deep neural network with local feature learning and recurrent regression. *Expert Systems with Applications*, 2017, 89: 66-80.

[11] FENG, Zhen-Hua, et al. Dynamic attention-controlled cascaded shape regression exploiting training data augmentation and fuzzy-set sample weighting. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017. p. 2481-2490.

[12] CAO, Xudong, et al. Face alignment by explicit shape regression. *International journal of computer vision*, 2014, 107: 177-190.

[13] WU, Yue; JI, Qiang. Shape augmented regression method for face alignment. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2015. p. 26-32.

[14] WANG, Xinyao; BO, Liefeng; FUXIN, Li. Adaptive wing loss for robust face alignment via heatmap regression. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019. p. 6971-6981.

[15] CHANDRAN, Prashanth, et al. Attention-driven cropping for very high resolution facial landmark detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020. p. 5861-5870.

[16] HINTON, Geoffrey; VINYALS, Oriol; DEAN, Jeff. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[17]WANG, Xiaojie, et al. Kdgan: Knowledge distillation with generative adversarial networks. *Advances in neural information processing systems*, 2018, 31.

[18]FINOGEEV, E., et al. KNOWLEDGE DISTILLATION USING GANS FOR FAST OBJECT DETECTION. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, 2020, 43.

[19]ZHOU, Zaida, et al. Channel distillation: Channel-wise attention for knowledge distillation. *arXiv preprint arXiv:2006.01683*, 2020.

[20] YOU, Shan, et al. Learning from multiple teacher networks. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2017. p. 1285-1294.

[21] SZEGEDY, Christian, et al. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[22] GOODFELLOW, Ian J.; SHLENS, Jonathon; SZEGEDY, Christian. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[23] KURAKIN, Alexey; GOODFELLOW, Ian; BENGIO, Samy. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

[24] YANG, Wei, et al. 3d human pose estimation in the wild by adversarial learning. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018. p. 5255-5264.

[25] CHEN, Guobin, et al. Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems*, 2017, 30.

[26] FARD, Ali Pourramezan; MAHOOR, Mohammad H. Facial landmark points detection using knowledge distillation-based neural networks. *Computer Vision and Image Understanding*, 2022, 215: 103316.

[27] KHABARLAK, Kostiantyn; KORIASHKINA, Larysa. Fast facial landmark detection and applications: A survey. *arXiv preprint arXiv:2101.10808*, 2021.

[28] BURGOS-ARTIZZU, Xavier P.; PERONA, Pietro; DOLLÁR, Piotr. Robust face landmark estimation under occlusion. In: *Proceedings of the IEEE international conference on computer vision*. 2013. p. 1513-1520.

[29] SAGONAS, Christos, et al. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: *Proceedings of the IEEE international conference on computer vision workshops*. 2013. p. 397-403.