ANALISIS EKSPRESI WAJAH BERDASARKAN KAEDAH RANGKAIAN NEURAL KONVOLUSIONAL

Jackson Kek Hai Wei
Kok Ven Jyn

Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Ekspresi wajah adalah cara semula jadi bagi manusia untuk menyampaikan perasaan serta emosi mereka. Ia juga merupakan isyarat bukan lisan yang penting untuk mencapai interaksi sosial yang berkesan. Otak manusia mempunyai mekanisme konstruktif yang membenarkan manusia untuk mengenali ekspresi wajah yang berbeza dengan serta-merta, tetapi ini merupakan satu tugasan yang mencabar untuk mesin. Hal ini demikian kerana terdapat pelbagai masalah yang perlu diatasi seperti masalah oklusi, pencahayaan, variasi manifestasi ekspresi wajah oleh individu yang berlainan dan lain-lain. Oleh itu, projek ini bertujuan untuk membangunkan satu model pembelajaran mendalam yang baru untuk klasifikasi ekspresi wajah yang memfokuskan kepada masalah variasi manifestasi ekspresi wajah oleh individu yang berlainan. Dalam projek ini, kami mempersembahkan satu seni bina rangkaian neural konvolusional baru yang bernama *Enhanced Attention Residual Network* (EAResNet) untuk mengatasi masalah tersebut. EAResNet ini secara khususnya merupakan gabungan dua mekanisme perhatian iaitu *Spatial Transformer Network* (STN) dan *Convolution Block Attention Module* (CBAM) dengan ResNet sebagai tulang belakang model. Penerapan kedua-dua mekanisme perhatian dalam seni bina ini adalah bertujuan untuk meningkatkan keupayaan model dalam memberi perhatian kepada bahagian-bahagian penting wajah untuk menentukan ekspresi wajah seseorang. Eksperimen yang lanjut juga dilakukan menggunakan set data penanda aras awam Facial Expression Recognition- 2013 (FER-2013) dan berjaya mencapai hasil ketepatan 70.36% yang setanding dengan kaedah yang sedia ada.

1 PENGENALAN

Keupayaan untuk memahami emosi manusia sangat penting untuk mewujudkan interaksi sosial yang positif dalam kehidupan seharian kita. Ekspresi wajah adalah salah satu isyarat bukan lisan yang membolehkan manusia untuk menyampaikan perasaan ataupun emosi mereka dalam komunikasi sosial (Adolphs R,1999). Kebanyakan masa, manusia dapat menyampaikan emosi mereka melalui ekspresi wajah dengan pergerakan wajah seperti pergerakan kerning, hidung dan bibir.

Walaupun otak manusia mempunyai mekanisme konstruktif yang membolehkan kita mengenali ekspresi wajah dengan serta-merta tanpa sebarang kesukaran, namun pengecaman ekspresi wajah yang tepat masih merupakan tugasan yang mencabar kepada mesin. Dalam cara tradisional, klasifikasi ekspresi wajah dapat dipecahkan kepada tiga langkah (Shan Li and Weihong Deng, 2018). Pertama, melakukan pengesanan muka daripada gambar, kedua, ekstrak beberapa ciri dari gambar menggunakan teknik seperti histogram berorientasikan kecerunan (HOG), Transformasi ciri invarian skala (SIFT), corak binari tempatan (LBP) (Yassin Kortli et al 2019) dan akhir sekali, ciri-ciri yang diekstrak dari gambar digunakan untuk melatih model pembelajaran mesin seperti logistik regresi, SVM, hutan rawak dan lain-lain. Akhirnya, model tersebut akan digunakan untuk meramal ekspresi wajah kepada 7 kategori emosi utama seperti gembira, sedih, marah, terkejut, ketakutan, neutral dan jijik. Pendekatan tradisional ini dapat berfungsi dengan baik di bawah set data yang biasa tetapi mula menunjukkan had mereka dengan set data yang lebih mencabar. Di samping itu, pendekatan ini memerlukan campur tangan manusia untuk mengeluarkan standard yang terpakai untuk mendapatkan ciri yang berguna.

Kebangkitan teknik pembelajaran mendalam telah mendapat peningkatan prestasi yang positif berbanding dengan pendekatan pembelajaran mesin tradisional dari segi prestasi dan ketepatan dalam tugas pengelasan ekspresi wajah. Pembelajaran mendalam adalah pendekatan baru dalam bidang kecerdasan buatan yang diilhamkan oleh struktur dan fungsi otak manusia yang dikenali sebagai rangkaian neural buatan (J. J. Hopfield, 1982). Sejak itu, pelbagai jenis seni bina rangkaian neural telah diterokai dan dikembangkan untuk meningkatkan prestasi mesin dalam klasifikasi ekspresi wajah (Shan Li, Weihong Deng, 2018). Rangkaian neural mendalam dapat menangkap secara automatik kedua-dua ciri yang bertahap rendah dan tinggi dari gambar wajah tanpa memerlukan campur tangan manusia berbanding dengan pendekatan tradisional.

Walau bagaimanapun, kebanyakan pendekatan ini masih menunjukkan kekurangan kemampuan model untuk membuat generalisasi di mana model tersebut menunjukkan prestasi yang baik di bawah set data gambar yang biasa tetapi prestasi merosot dalam set data yang lebih mencabar. Hal ini demikian kerana dalam set data biasa, kebanyakan

ekspresi wajah yang ditunjukkan adalah dalam aplitud yang sama. Manakala dalam set data yang lebih mencabar, manifestasi wajah oleh orang yang berlainan mungkin menunjukkan ekspresi wajah yang berlainan dalam emosi yang sama. Sebagai contoh, individual yang lebih pemalu berkemungkinan menunjukkan ekspresi wajah yang tidak ketara berbanding dengan individual yang lebih suka membesar-besarkan ekspresi wajah mereka.

2 PENYATAAN MASALAH

Dalam beberapa dekad yang lalu, klasifikasi ekspresi wajah telah menjadi salah satu bidang penyelidikan aktif dalam pelbagai domain. Mengklasifikasikan ekspresi wajah dengan tepat menjadi satu tugasan yang mencabar terhadap mesin terutamanya dalam set data di mana terdapat variasi ekspresi wajah yang besar oleh individu yang berlainan dalam kategori ekspresi yang sama. Salah satu contoh dalam masalah variasi ekspresi wajah adalah seperti dalam satu emosi yang sama, individual yang berbeza mempunyai manifestasi ekspresi wajah yang berbeza. Sebagai contoh, sebilangan individu akan menunjukkan ekspresi wajah yang jelas, manakala sebilangan individu menunjukkan emosi mereka hanya dengan ekspresi secara separa ataupun mikro. Pendekatan yang sedia ada untuk menyelesaikan masalah tersebut adalah menggunakan teknik fitur kerajinan tangan atau rangkaian neural mendalam di mana kedua-dua ciri tahap rendah sehingga tinggi diekstrak dari gambar dan dimasukkan ke dalam pengklasifikasi untuk latihan. Namun, disebabkan pendekatan ini akan mendapatkan kebanyakkan ciri dari wajah tanpa sebarang tapisan, jadi ini akan menyebabkan hasil bilangan parameter yang besar untuk dipelajari oleh model. Ciri-ciri ini juga berkemungkinan merangkumi ciri-ciri yang tidak penting seperti telinga dan rambut yang tidak memainkan peranan dalam klasifikasi ekspresi. Oleh itu, keadaan ini akan menyebabkan situasi *overfitting* dalam klasifikasi ekspresi wajah. Sebahagian besar pendekatan pada masa kini hanya dapat menunjukkan prestasi yang baik dalam set data yang tidak mempunyai masalah variasi ekspresi wajah berlainan yang besar. Oleh itu, projek ini akan dilaksanakan dengan tujuan untuk mengatasi masalah ini.

3 OBJEKTIF KAJIAN

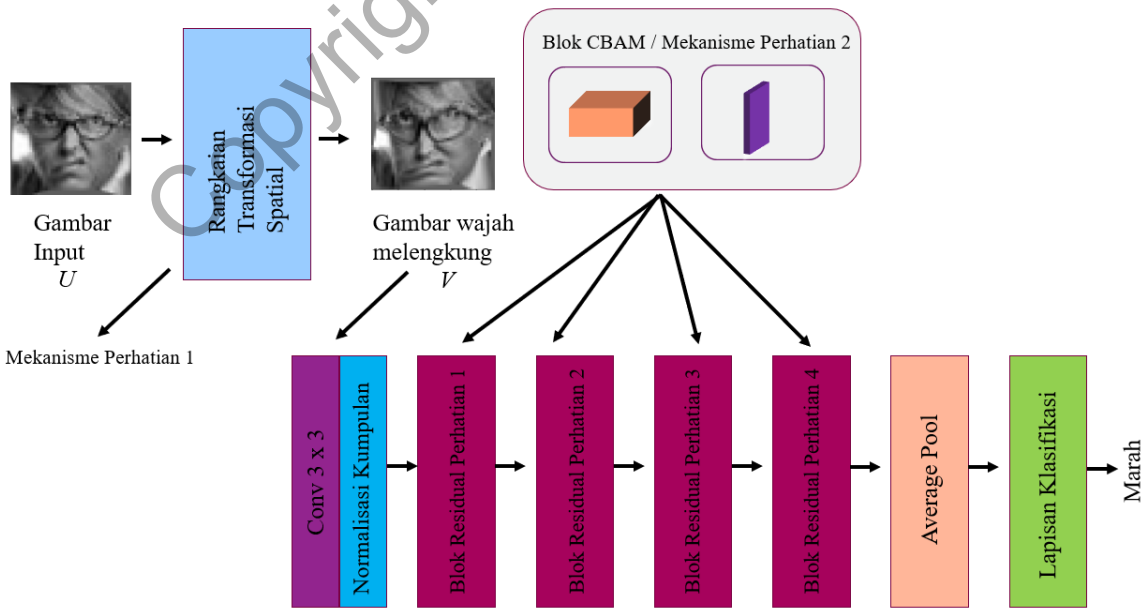
Objektif untuk projek ini adalah untuk mengembangkan model klasifikasi ekspresi wajah yang baru dan kuat untuk mengklasifikasikan ekspresi wajah kepada emosi yang berbeza, seperti gembira, sedih, marah, kejutan, ketakutan, jijik, dan berkecuali. Lebih khusus lagi, projek ini bertujuan untuk:

i. Mencadang seni bina rangkaian neural konvolusional yang baru dengan penerapan mekanisme perhatian untuk mempelajari ciri-ciri penting dan mengklasifikasikan setiap ekspresi wajah ke dalam kategori yang berbeza.

ii. Menggambarkan bahagian-bahagian penting dari ekspresi wajah yang mempunyai pengaruh penting terhadap hasil klasifikasi.

iii. Mengesahkan model yang dicadangkan dengan set data penanda aras wajah berdasarkan kadar ketepatan.

4 METOD KAJIAN



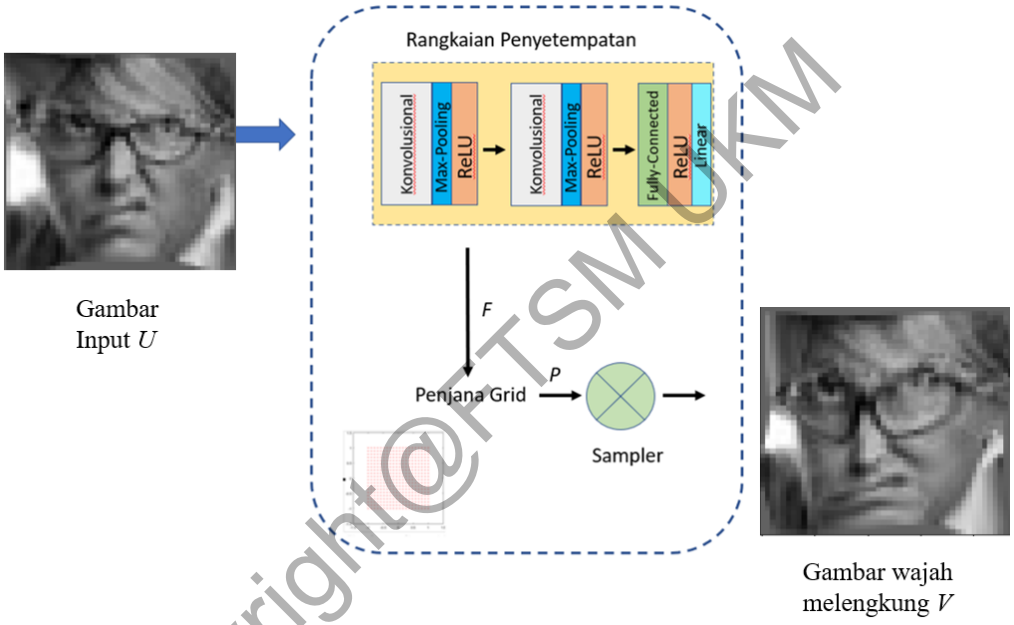
Rajah 4.1 Seni bina EAResNet-18 untuk klasifikasi ekspresi wajah secara keseluruhan

Persepsi visual manusia mempunyai struktur perhatian yang membolehkan kita menumpukan perhatian pada bahagian-bahagian penting wajah untuk menentukan ekspresi wajah seseorang. Berlainan dengan manusia pula, rangkaian pembelajaran mendalam tradisional memerlukan penggabungan mekanisme perhatian visual secara eksplisit ke dalam rangkaian untuk memfokuskan bahagian-bahagian penting dalam wajah. Oleh itu, seni bina yang dicadangkan ini akan mengabungkan mekanisme perhatian yang sebahagian besarnya diilhami oleh sistem persepsi visual manusia ke dalam rangkaian saraf konvolusional untuk mempelajari ciri-ciri penting dari pelbagai bahagian wajah. Disebabkan penerapan mekanisme perhatian ke dalam seni bina ResNet, model yang dicadangkan ini dinamakan sebagai "Enhanced Attention Residual Network" ataupun boleh diringkaskan sebagai EAResNet. Secara khususnya, EAResNet ini adalah seperti yang ditunjukkan dalam Rajah 4.1 dan merupakan gabungan antara seni bina model ResNet (Kaiming He et al, 2015) dengan dua komponen mekanisme perhatian iaitu Rangkaian Transformasi Spatial (*Spatial Transformation Network*) (Max Jaderberg et al. 2015) sebagai mekasnime perhatian pertama dengan modul CBAM (*Convolutional Block Attention Module*) (Sanghyun Woo et al. 2018) yang dimasukkan ke dalam setiap blok ResNet untuk membentuk blok residual perhatian ataupun mekanisme perhatian kedua.

Proses dalam EAResNet adalah bermula dengan pengambilan gambar wajah gambar wajah input $U \in \mathbb{R}^{H \times \mathrm{W} \times \mathrm{C}}$ dengan ketinggian H, lebar W dan bilangan saluran warna C sebagai input ke dalam rangkaian transformasi spatial untuk melakukan transformasi terhadap gambar input supaya dapat menghasilkan gambar wajah melengkung $V \in \mathbb{R}^{H \times \mathrm{W} \times \mathrm{C}}$. Gambar wajah melengkung ini merupakan pengubahsuaian terhadap gambar input untuk menonjolkan bahagian-bahagian wajah yang penting seperti mata, hidung dan mulut kepada klasifikasi ekspresi wajah. Gambar wajah melengkung tersebut akan dimasukkan ke dalam lapisan konvolusional bersama dengan normalisasi kumpulan untuk mendapatkan pewakilan ciri (*feature representation*) dari wajah. Seterusnya, pewakilan ciri tersebut akan dimasukkan ke dalam blok-blok residual perhatian untuk mengekstrak ciri-ciri yang penting dalam penentuan ekspresi wajah. Peta ciri yang diekstrak akan dimasukkan ke dalam lapisan Average Pool untuk mengurangkan dimensi peta ciri tersebut. Akhirnya, peta ciri yang dikurangkan dimensi akan diambil oleh lapisan

fully-connected untuk menghasilkan vektor $P = \begin{bmatrix} p_1 \\ p_2 \\ . \\ . \\ p_E \end{bmatrix}$ di mana E adalah bilangan kategori ekspresi dan setiap elemen p mewakili keberangkalian gambar input untuk setiap kategori ekspresi.

4.1 Rangkaian Transformasi Spatial (STN)



Rajah 4.2 Rangkaian transformasi spatial. Gambar input akan dimasukkan ke dalam rangkaian penyetempatan untuk mendapatkan titik *F* yang dapat menentukan bagaimana gambar akan digubah. Titik *F* akan diambil oleh penjana grid untuk menghasilkan grid yang dapat diterapkan ke dalam gambar input oleh Sampler untuk menghasilkan gambar wajah melengkung *V*.

Rangkaian transformasi spatial ini merupakan mekanisme perhatian yang pertama dalam EAResNet. Rangakaian ini bertujuan untuk mengurangkan kesan manifestasi ekspresi wajah yang berbeza oleh individu yang berlainan dengan melakukan transformasi terhadap gambar wajah untuk mendapatkan gambar wajah melengkung yang dapat menonjolkan bahagian paling penting dalam wajah manusia. Diberikan gambar input *U*, rangkaian ini akan melakukan transformasi dan menghasilkan gambar wajah melengkung V yang bersaiz

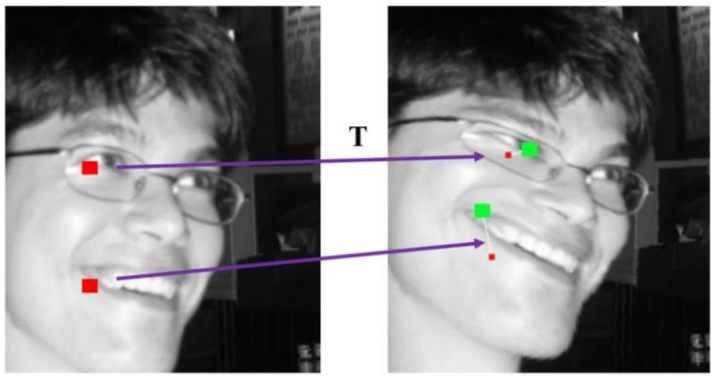
sama dengan input. STN ini mempunyai tiga komponen yang utama, iaitu Rangkaian Penyetempatan, Penjana grid dan Sampler.

Rangkaian Penyetempatan

Rangkaian penyetempatan dalam STN ini adalah rangkaian neural konvolusional biasa yang terdiri daripada dua lapisan konvolusional di mana setiap lapisan diikuti dengan fungsi pengaktifan maksimum (Max-Pooling) dan pengaktifan ReLU, dan dua lapisan fully-connected. Matlamat utama rangkaian penyetempatan adalah untuk meramal dan menghasilkan bilangan K titik rujukan (*fiducial points*) daripada gambar input yang akan digunakan dalam transformasi TPS untuk diterapkan pada gambar input. Titik rujukan yang dihasilkan ini akan diwakili dengan $F = [f_1 \ldots . f_K] \in \mathbb{R}^K$, di mana setiap $f_k = [x_k, y_k]^T$ yang mengandungi koordinat K-th titik rujukan.

Penjana Grid

Penjana grid bertanggungjawab untuk meramalkan parameter yang diperlukan untuk transformasi TPS dan menghasilkan grid sampling. Dalam penjana grid ini, satu lagi set titik rujukan asas (*base fiducial* point) yang lain $F' = [f'_1 \ldots . f'_K] \in \mathbb{R}^K$ akan dtentukan. Tujuan untuk menentukan dua set titk rujukan adalah untuk melakukan transformasi terhadap gambar input dari titik rujukan F kepada titik rujukan asas F' melalui transformasi TPS pada akhir peringkat STN. Contoh gambaran kesan untuk penggunaan dua set titik rujukan dalam transformasi TPS pada peringkat akhir telah ditunjukkan dalam Rajah 4.3.



Rajah 4.3 Contoh gambaran kesan penggunaan dua set titik rujukan di mana K = 2 terhadap gambar wajah. Titik merah di sebelah kiri mewakili titik rujukan F dan titik hijau dibelah kanan mewakili titik rujukan asas F'. Garisan berwarna ungu mewakili transformasi TPS $\boldsymbol{T}$. Untuk setiap titk dalam F', $[x'_i, y'_i]$ dalam V, transformasi T akan mencari titik $[x_i, y_i]$ dari input U dan melakukan transformasi.

Secara ringkasnya, parameter untuk transformasi TPS $\boldsymbol{T}$ boleh diwakili dengan,

$$T = (\Delta_{F'}^{-1} \begin{bmatrix} F^T \\ 0^{3\ X\ 2} \end{bmatrix})^T \tag{1}$$

di mana $\Delta_{F'} \in \mathbb{R}^{(K+3)(K+3)}$ merupakan matriks

$$\Delta_{F'} = \begin{bmatrix} 1^{K-1} & F'^T & \boldsymbol{R} \\ 0 & 0 & 1^{1\ x\ K} \\ 0 & 0 & F' \end{bmatrix} \tag{2}$$

Manakala, elemen dalam i-th barisan dan j-th kolum dalam $\boldsymbol{R}$ di persamaan (2) adalah $r'_{i,j} = d_{i,j}^2 \ln(d_{i,j}^2)$, dan $d_{i,j}$ adalah jarak euclidean antara f'_i dengan f'_j.

Untuk grid P' pada output V pula, P' boleh diwakili dengan $P' = \{\mathbf{p}'_i\}_{i=1\ldots N}$, di mana $\mathbf{p}'_i = [x'_i, y'_i]^T$ adalah x dan y koordinat untuk i-th piksel dan N adalah jumlah piksel. Seperti yang ditunjukkan dalam Rajah 4.3, untuk setiap titik piksel $\mathbf{p}'_i$ dari V, transformasi akan dilakukan terhadap titik $\mathbf{p}_i = [x_i, y_i]$ yang sepadan seperti:

$$r'_{i,k} = d_{i,k}^2 \ln(d_{i,k}^2) \tag{3}$$

$$\hat{\mathbf{p}}_\mathbf{i} = [1, x'_i, y'_i, r'_{i,1}, \ldots., r'_{i,K}]^T, \tag{4}$$

$$\mathbf{P}_i = T\hat{\mathbf{p}}_i \tag{5}$$

di mana $d_{i,k}$ adalah jarak euclidean antara $\mathbf{p}'_i$ dengan k-th titik rujukan asas f'_i. Selepas setiap titik telah dilakukan transformasi, satu grid sampling $P = \{\mathbf{p}_i\}_{i=1\ldots N}$ akan dihasilkan.

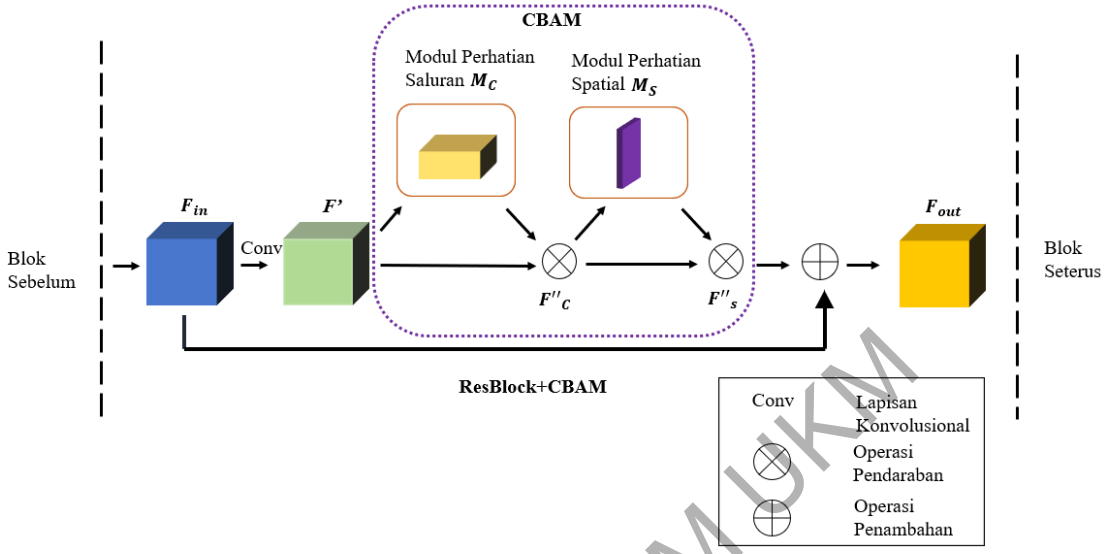
Sampler

Akhirnya, sampler ini akan mengambil grid sampling *P* yang dihasilkan untuk diterapkan kepada gambar input dengan menggunakan interpolasi bilinear untuk dan menghasilkan output peta ciri transformasi yang melengkung. Interpolasi bilinear yang digunakan oleh pensampel adalah $V = B(P, U)$ dimana B adalah interpolasi bilinear yang boleh dikembangkan kepada:

$$V_i^c = \sum_n^H \sum_m^W U_{nm}^c \max(0,1 - |x_i - m|) \max(0,1 - |y_i - n|) \tag{6}$$

di mana U'^c_i adalah nilai output untuk piksel *i* di lokasi (x_i, y_i) di saluran warna *C*. U_{nm}^c adalah nilai input di lokasi (*n*, *m*) di saluran warna *C*. Antara sebab untuk menggunakan interpolasi bilinear adalah untuk memetakan hasil grid pensampelan yang berkemungkinan dalam bentuk pecahan (*fractional* form) kepada nilai piksel integer yang sesuai. Akhirnya, gambar wajah melengkung *V* akan dihasilkan dan dimasukkan ke dalam satu lapisan konvolusional dengan saiz kernel 3x3 dahulu untuk mendapatkan pewakilan ciri (*feature representation*) yang umun. Pewakilan ciri ini seterusnya akan dimasukkan ke dalam blok residual perhatian untuk mengekstrak ciri-ciri yang penting.

4.2 Blok Residual Perhatian (Sanghyun Woo et al. 2018)



Rajah 4.4 Komponen dan proses dalam blok residual perhatian (ResBlock+CBAM).

Blok residual perhatian ini pula merupakan mekanisme perhatian dalam EAResNet yang kedua yang digunakan untuk meningkatkan lagi keupayaan model untuk memberi perhatian kepada ciri-ciri penting. Seperti yang ditunjukkan dalam Rajah 4.4, blok residual perhatian ini merupakan gabungan antara ResBlock dalam ResNet dengan blok CBAM yang mempunyai mekanisme perhatian untuk membolehkan model memberikan perhatian kepada ciri-ciri yang penting dan mengurangkan kesan akibat gangguan ciri-ciri yang kurang penting terhadap hasil klasifikasi. Setiap blok ini akan mengambil peta ciri F_{in} dari blok yang sebelumnya sebagai input dan mengekstrak ciri-ciri untuk membentuk peta ciri $F' \in \mathbb{R}^{C \times H \times W}$ yang mengandungi pewakilan ciri-ciri yang bertahap rendah dan tinggi daripada ekspresi wajah menggunakan lapisan konvolusional yang diwakili sebagai "Conv" dalam Rajah 4.4. Peta ciri F' ini akan dimasukkan ke dalam blok CBAM yang terdiri daripada Modul Perhatian Saluran, M_c dan Modul Perhatian Spatial, M_s untuk menghasilkan peta ciri yang dipertingkatkan perhatian (*attention-enhanced feature map*) F_{out} yang mengandungi ciri-ciri yang bermaklumat dan penting. Modul perhatian saluran, M_c ini mampu untuk menentukan bahagian "apa" yang hendak diberi perhatian oleh model

dan modul perhatian spatial dapat menentukan bahagian "mana" yang penting untuk hasil klasifikasi. Proses dalam blok ini boleh diringkaskan kepada operasi berikut:

$$F''_c = M_c(F') \otimes F' \quad \ldots(4.1)$$

$$F''_s = M_s(F''_c) \otimes F''_c$$

$$F_{out} = F''_s \oplus F_{in}$$

Di mana $\otimes$ dan $\oplus$ mewakili operasi pendaraban dan penambahan secara unsur (*element-wise*). Pertama sekali, peta ciri F' yang dihasilkan selepas lapisan konvolusional akan dimasukkan ke dalam modul perhatian saluran yang diwakili oleh M_c dan menghasilkan peta saluran perhatian (*channel attention map*) yang akan didarab semula dengan F' untuk mendapatkan peta ciri diperhalusi saluran (*channel-refined feature map*) F''_c. Output F''_c pula akan dimasukkan ke dalam modul perhatian spatial M_s dan menghasilkan peta spatial perhatian (*spatial attention map*) yang akan diterapkan semula ke F''_c dengan operasi pendaraban untuk mendapatkan peta ciri diperhalusi secara spatial (*spatial-refined feature map*) F''_s. Akhirnya, peta ciri F''_s akan ditambah dengan peta ciri input F_{in} seperti yang ditunjukkan dalam Rajah 4.4 untuk menghasilkan peta ciri F_{out} yang mengandungi ciri-ciri yang penting pada tahap tertentu berdasarkan lapisan blok residual perhatian.

4.3 Fungsi Kerugian (Loss Function)

Fungsi kerugian adalah digunakan untuk menilai perbezaan antara keberangkalian setiap kategori ekspresi yang dihasilkan oleh model dengan kategori ekspresi yang sebenar. Untuk projek ini, fungsi kerugian yang digunakan dalam seni bina ini adalah cross-entropy teratur (*regularized cross*-entropy) yang dilambangkan sebagai $L_{overall}$ untuk klasifikasi kerugian. . $L_{overall}$ hanyalah penjumlahan cross_ entropy $L_{classifier}$ dengan istilah teratur (*regularized term*) yang merupakan norma L_2 pemberat dari dua lapisan fully-connected yang terakhir.

$$L_{overall} = L_{classifier} + \lambda\left|\left|w_{(fc)}\right|\right|_2^2 \quad \ldots(4.2)$$

$$L_{classifier} = -(ylog(P) + (1-y)\log(1-P)) \quad \ldots(4.3)$$

di mana penggal pertama $L_{classifier}$ mewakili persamaan untuk cross-entropy (Zhilu dan Mert, 2018) dan istilah kedua yang ditambahkan adalah norma L_2 atau jumlah kuadrat (*sum of square*) dari kesemua berat dalam dua lapisan fully-connected yang terakhir. Dalam $L_{classifier}$, y merupakan label yang betul daripada set data dan *P* mewakili output ramalan daripada model ini. Nilai parameter regularisasi, λ ini akan ditala berdasarkan hasil ketepatan pada set pengesahan. Antara sebab untuk menambahkan istilah tambahan ke dalam fungsi kerugian adalah untuk mengurangkan kemungkinan *overfitting* model yang disebabkan bilangan gambar yang kecil dalam set data. Keseluruhan model akan dilatih dengan meminimumkan fungsi kerugian menggunakan penurunan kecenderungan stokastik (*stochastic gradient descent*) atau lebih khusus lagi pengoptimum Adam.

4.4 Tetapan Eksperimen

Eksperimen ini adalah menumpu kepada set data FER-2013 yang terdiri daripada 28,709 gambar wajah untuk proses latihan dan 3,589 masing-masing untuk set data pengesahan dan ujian. Model ini dilatih dalam kumpulan bersaiz 256 secara keseluruhan dengan menggunakan 250 epok. Untuk mengurangkan daya pengiraan dan masa yang diperlukan, EAResNet ini akan menggunakan ResNet-18 sebagai tulang belakang (*backbone*). Untuk pengoptimuman, pengoptimum Adam telah digunakan dengan kadar pembelajaran (*learning* rate) sebanyak 0.001 dan akan dikurangkan pada setiap 50 epok. berserta dengan penurunan berat (*weight decay*) sebanyak 1e-4. Untuk transformasi TPS dalam modul perhatian pula, beberapa set titik rujukan telah digunakan dan 5 titik rujukan menunjukkan prestasi yang paling optimum. Keseluruhan proses ini dijalankan di persekitaran maya dalam kaggle kernel yang menggunakan GPU Nvidia K80 seperti yang dibincangkan dalam bab awal. Disebabkan set data FER-2013 mempunyai ketidakseimbangan bilangan gambar dalam sesetengah kategori, teknik pembesaran data telah digunakan. Teknik pembesaran data yang digunakan adalah berdasarkan kepada Christopher Pramerdorfer dan Martin

Kampel, 2016 yang menggunakan teknik seperti permotongan saiz rawak dan teknik membalik mendatar secara rawak.

5 HASIL KAJIAN

Menurut Amil Khanzada et al 2020, prestasi tahap manusia dalam mengklasifikasi ekspresi wajah dalam FER-2013 adalah 65±5%. Namun begitu, model yang dicadangkan telah mencapai 70.36% yang setanding dengan prestasi tahap manusia tanpa menggunakan set data latihan tambahan. Disebabkan had dalam perkakasan, tulang belakang yang digunakan dalam model ini adalah ResNet-18 yang hanya mempunyai 18 lapisan CNN. Selain itu, perbandingan hasil kajian oleh model yang dicadangkan dengan beberapa kajian yang sebelumnya telah ditunjukkan dalam Jadual 6.1.

Jadual 5.1 Perbandingan model yang dicadangkan dengan kaedah sebelumnya

Kaedah	Ketepatan
Bag of Words (Ionescu et al. 2013)	67.4%
GoogleNet (Giannopoulos et al. 2018)	65.2%
VGG+SVM (Georgescu et al. 2018)	66.31%
ResNet (H. Ma and T. Celik, 2019)	66.51%
Deep-Emotion (Shervin Minaee & Amirali Abdolrashidi, 2019)	70.02%
EAResNet-18 (Model yang dicadangkan)	**70.36%**

Dalam Jadual 5.1, beberapa kaedah yang sedia ada telah dipilih untuk membuat perbandingan dengan EAResNet yang dicadangkan dan disusunkan mengikut tahun. Kaedah-kaedah seperti GoogleNet, VGG+SVM, ResNet dan Deep-Emotion merupakan kaedah yang menggunakan seni bina CNN yang berlainan manakala Bag of Words merupakan kaedah pembelajaran mesin yang tradisional. Kebanyakkan kaedah ini dipilih untuk membuat perbandingan adalah disebabkan kaedah-kaedah tersebut mempunyai tetapan eksperimen yang serupa dan tidak menggunakan set data latihan tambahan. Perbandingan dalam Jadual 5.1 ini menunjukkan bahawa EAResNet-18 telah berjaya mendapatkan keputusan yang setanding dengan kaedah yang sedia ada dengan pencapaian ketepatan sebanyak 70.36% dalam klasifikasi ekspresi wajah FER-2013. ResNet (H. Ma and T. Celik, 2019) yang ditunjukkan dalam jadual perbandingan adalah seni bina ResNet-18 yang mempunyai 18 lapisan seperti yang digunakan dalam EAResNet dan mempunyai ketepatan sebanyak 66.51%. Ini menunjukkan bahawa penerapan mekanisme perhatian ke dalam ResNet barjaya meningkatkan prestasi model klasifikasi ekspresi wajah sebanyak 3.85%. Manakala, EAResNet-18 juga berjaya memperoleh prestasi yang lebih tinggi berbanding dengan Deep-Emotion yang hanya mencapai hasil ketepatan sebanyak 70.02% dengan penerapan mekanisme perhatian tunggal iaitu STN ke dalam seni bina CNN. Perbandingan ini telah membuktikan bahawa EAResNet yang menerapkan mekanisme perhatian yang kedua iaitu CBAM ke dalam seni binajuga mampu meningkatkan lagi prestasi model.

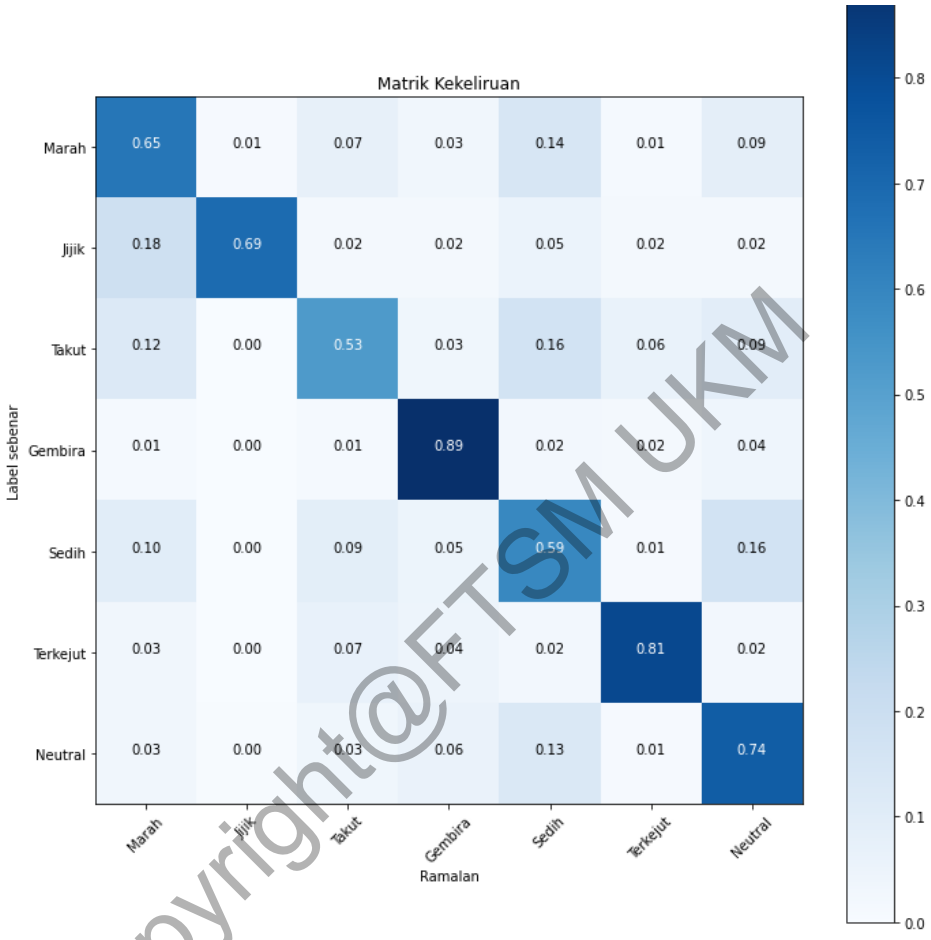
Seperti yang dibincangkan dalam awal bab ini, FER-2013 merupakan set data yang lebih mencabar berbanding dengan set data ekspresi wajah yang lain. Antara cabaran utama set data ini adalah ketidakseimbangan untuk setiap kategori ekspresi wajah. Kategori ekspresi wajah seperti gembira dan neutral mempunyai bilangan berbanding dengan kategori yang lain. Seperti yang ditunjukkan dalam Jadual 5.2, model yang dicadangkan dapat mencapai ketepatan yang lebih tinggi dalam kategori gembira (89%) dengan kategori terkejut (81%). Manakala, prestasi model telah merosot dalam proses mengklasifikasi ekspresi wajah dalam kategori takut (53%) dan sedih (59%).

Kategori	Ketepatan	Jumlah gambar
Marah	65%	491
Jijik	69%	55
Takut	53%	528
Gembira	89%	879
Sedih	59%	594
Terkejut	81%	416
Neutral	74%%	626

Jadual 5.2 Ketepatan model untuk setiap katgori dalam set data ujian

Oleh itu, matriks kekeliruan telah ditunjukkan dalam Rajah 5.1 untuk menganalisis hasil ketepatan model dalam setiap kategori secara terperinci. Label di sebelah kiri matrik kekeliruan merupakan label sebenar untuk ekspresi wajah tersebut dan label di bawah merupakan hasil ramalan model. Setiap nilai dalam barisan pepepenjuru merupakan peratusan ramalan yang betul untuk kategori tersebut. Sebagai contoh, nilai 0.89 di bahagian pepenjuru untuk kategori gembira adalah menunjukkan bahawa EAResNet berjaya meramalkan 89% daripada kategori gembira dengan betul. Untuk nilai selain daripada pepenjuru pula, ia merupakan peratusan yang salah diklasifikasikan oleh model ke kategori yang lain. Sebagai contoh, nilai dalam barisan pertama dan lajur kedua iaitu 0.01 bermaksud terdapat 1% daripada ekspresi wajah dalam kategori marah telah salah diklasifikasikan ke dalam kategori jijik. Manakala, terdapat 14% ekspresi wajah dalam kategori marah telah salah diklasifikasikan ke dalam kategori sedih. Melalui matriks kekeliruan, kebanyakkan masa model yang dicadangkan sukar untuk membezakan anatara ekspresi wajah sedih dengan ekspresi wajah yang neutral. Selain itu, kekeliruan juga

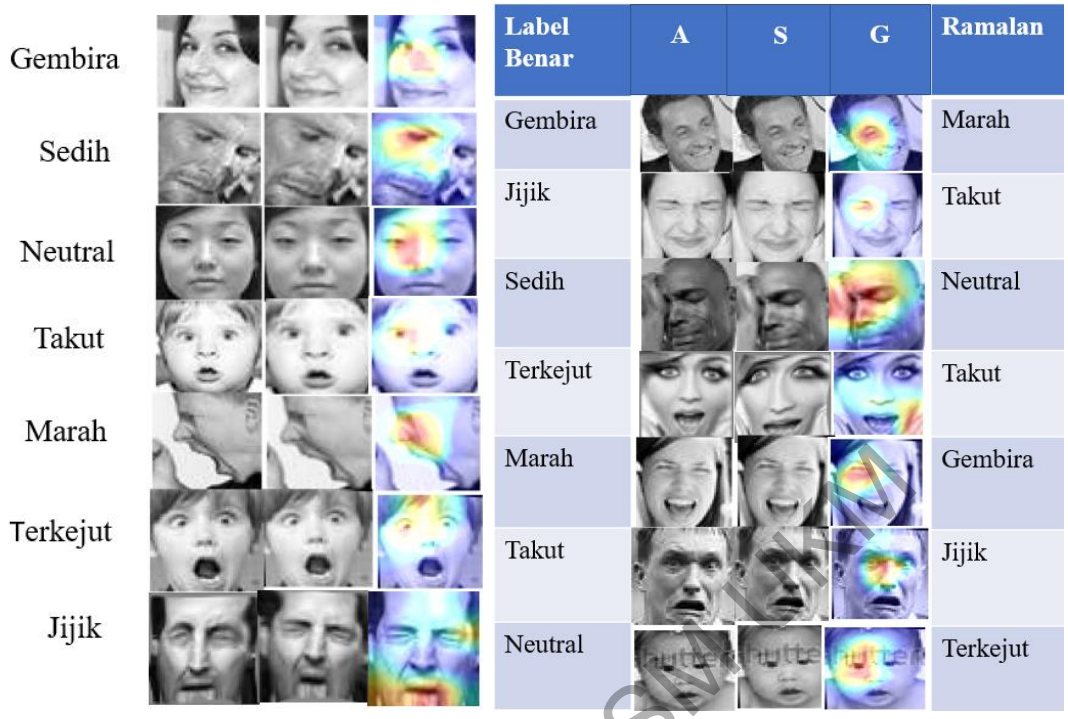
sering menimbul dalam kategori ekspresi wajah seperti kemarahan, kesedihan dan ketakutan kerana ketiga-tiga kategori ini sering menunjukkan gerakkan wajah yang serupa.



Rajah 5.1 Matriks kekeliruan untuk FER-2013

5.1 Visualisasi Model

Dalam bahagian ini, penggunaan Grad-Cam yang diilham oleh Ramprasaath R. Selvaraju et al pada 2019 telah digunakan sebagai kaedah untuk mengambarkan bahagian-bahagian penting kepada model dalam penentuan ekspresi wajah. Kaedah ini dapat digunakan untuk mengambarkan peta pengaktifan (*activation map*) oleh mana-mana lapisan konvolusional. Untuk bahagian ini, peta pengaktifan dalam lapisan konvolusional yang terakhir telah dipilih sebagai visualisasi untuk model.



Rajah 5.2 Kiri: Ramalan model yang betul bersama dengan hasil transformasi dalam STN di bahagian tengah dan visualisasi Grad-Cam. Kanan: Perbandingan antara label sebenar dengan ramalan model. A mewakili gambar wajah asal, S mewakili gambar wajah melengkung dan G mewakili visualisasi Grad-Cam. Peta warna jet digunakan dalam Grad-Cam di mana warna merah mewakili bahagian yang mempunyai kesan paling tinggi terhadap hasil klasifikasi dan warna biru adalah paling rendah.

Rajah 5.2 menunjukkan perbandingan visualisasi model dalam ramalan kategori ekspresi yang betul dengan ramalan model yang salah. Dalam Rajah 5.2 ini, model yang dicadangkan adalah mempunyai keupayaan untuk memfokuskan kepada bahagian-bahagian yang penting dari ekspresi wajah. Sebagai contoh, model ini memfokuskan kepada bahagian mata untuk kategori neutral dan bahagian mulut semasa menentukan ekspresi wajah jijik dalam ramalan betul. Namun demikian, EAResNet-18 ini juga menunjukkan kes gagal seperti yang ditunjukkan di sebelah kanan Rajah 5.2. Untuk kategori terkejut, EAResNet-18 telah menunjukkan kehilangan keupayaan dalam memberi perhatian kepada bahagian-bahagian penting wajah dengan memfokuskan kepada bahagian yang tidak relevan. Selain itu, visualisasi ini juga jelas menunjukkan bahawa model yang dicadangkan ini hanya dapat memfokuskan kepada satu bahagian yang tertentu sahaja dan

tidak mempunyai keupayaan untuk memberi perhatian kepada pelbagai bahagian secara serentak.

6 KESIMPULAN

Project ini bertujuan untuk memajukan sistem klasifikasi ekspresi wajah manusia dengan membangunkan satu seni bina CNN yang baru – EAResNet untuk menangani masalah variasi ekspresi wajah oleh individu yang berlainan. EAResNet ini merupakan peningkatan kepada model perhatian tunggal dalam kajian-kajian yang sedia ada. EAResNet ini mempunyai dua mekanisme perhatian utama iaitu STN dan CBAM untuk meningkatkan keupayaan model dalam memberi perhatian kepada bahagian-bahagian wajah yang penting dalam klasifikasi ekspresi wajah. Eksperimen telah dijalankan dengan menggunakan set data penanda aras FER-2013 dan hasil kajian telah menunjukkan bahawa model yang dicadangkan mampu mencapai tahap manusia dan mempunyai prestasi yang setanding dengan kaedah yang sedia ada dalam klasifikasi ekspresi wajah. Penambahbaikan yang boleh dilakukan pada masa hadapan adalah dengan menggunakan tulang belakang model yang lebih mendalam seperti ResNet-50 ataupun ResNet-101. Selain itu, gambar wajah juga boleh dibesarkan kepada saiz standard iaitu 224 x 244 dan menjalankan proses latihan dalam pekakasan yang mempunyai daya pengiraan yang lebih kuat. Selain itu, bilangan set data juga boleh ditambah untuk mengelakkan masalah kategori kelas ekspresi wajah yang tidak seimbang. Dari segi model pula, teknik transformasi dalam STN yang berbeza boleh digunakan seperti transformasi projektif ataupun pegabungan transformasi seperti transformasi affin dan TPS.

7 RUJUKAN

Amil Khanzada, Charles Bai, Ferhat Turker Celepcikay, 2020. Facial Expression Recognition with Deep Learning Improving on the State of the Art and Applying to the Real World.

J. Adolphs R. 1999. Social cognition and the human brain. Trends Cogn Sci 3(12):1-2.

J Hopfield 1982. Neural networks and physical systems with emergent collective computational abilities Proceedings of the National Academy of Sciences Apr 1982, 79 (8) 2554-2558

Shan Li dan Weihong Deng 2018. Deep facial expression recognition: A survey. IEEE Transactions on Affective Computing. 2020.

Yassin Kortli, Maher Jridi b, Ayman Al Falou b, dan Mohamed Atr, 2018. A comparative study of CFs, LBP, HOG, SIFT, SURF, and BRIEF for security and face recognition.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun 2016. Deep residual learning for image recognition.In Proceedings of the IEEE conference on computer vision and pattern recognition 770-778

Minaee, Shervin, Mehdi Minaei, and Amirali Abdolrashidi. 2021. "Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network" Sensors 21, no. 9: 3046.

Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al, 2013. Challenges in representation learning: A report on three machine learning contests. In International Conference on Neural Information Processing, 117–124.

Jaderberg, Max, Karen Simonyan, and Andrew Zisserman, 2015. Spatial transformer networks. Advances in neural information processing systems.

Giannopoulos, Panagiotis, Isidoros Perikos, and Ioannis Hatzilygeroudis 2018. ”Deep Learning Approaches for Facial Emotion Recognition: A Case Study on FER-2013.” Advances in Hybridization of Intelligent Methods. Springer, Cham, 1-16.

Georgescu, Mariana-Iuliana, Radu Tudor Ionescu, and Marius Popescu. ”Local Learning with Deep and Handcrafted Features for Facial Expression Recognition.” arXiv preprint arXiv:1804.10892, 2018.

Ionescu, Radu Tudor, Marius Popescu, and Cristian Grozea. ”Local learning to improve bag of visual words model for facial expression recognition.” Workshop on challenges in representation learning, ICML. 2013.

H. Ma and T. Celik, “Fer-net: facial expression recognition using densely connected convolutional network,” Electronics Letters, vol. 55, no. 4, pp. 184–186, 2019.

R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618-626, doi: 10.1109/ICCV.2017.74.

Shan Li, Weihong Deng, and JunPing Du, 2017. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. IEEE Conference on Computer Vision and Pattern Recognition. 2852–2861

Zhilu Zhang dan Mert R. Sabuncu 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems. 8792–8802.

George Bebisa, Michael Georgiopoulosb, Nielsda Vitoria Loboc dan MubarakShahc, 1999. Pattern Recognition 32, Issue 10,1783-1799.