REBIFFACE: PENGESAN WAJAH DWI-GABUNGAN SISA UNTUK PENGESANAN WAJAH DALAM GAMBAR HIMPUNAN ORANG RAMAI

Edmund Ngu Jan Piew
Dr. Kok Ven Jyn

Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Mengesan wajah kecil dalam himpunan orang ramai yang padat merupakan salah satu masalah yang mencabar dalam domain penglihatan komputer. Dalam himpunan orang ramai yang padat, bilangan individu yang banyak menyebabkan oklusi teruk antara individu dan skala wajah yang berbeza, membawa masalah maklumat semantik konteks (terdiri daripada lokasi, tekstur, dan skala spasial) yang tidak mencukupi dalam ciri-ciri yang diekstrak untuk pengesanan wajah yang tepat. Walaupun terdapat banyak kajian mengenai pengesanan wajah kecil, penyelidikan ini masih mempunyai banyak ruang untuk berkembang dengan kemajuan pembelajaran mendalam dalam penglihatan komputer. Dalam projek ini, kami mempersembahkan satu pengesan wajah satu-tahap yang bernama RebifFace untuk mencapai prestasi tinggi dengan mengambil kesempatan daripada pembelajaran pelbagai tugas lima mercu tanda muka yang diawasi secara ekstra. Secara khususnya, RebifFace terdiri daripada dua modul: Residual Bi-Fusion (ReBiF) Feature Pyramid (FP) dan Receptive Field Enhancement (RFE). Untuk mengekalkan ketepatan pengesanan wajah bersaiz besar dan kecil, ReBiF menggabungkan kedua-dua ciri dalam dan cetek dalam FP dengan cara dua arah (jalur atas-bawah dan bawah-atas) untuk memelihara maklumat wajah yang tepat dari kesan pergeseran pooling. Selain itu, RFE diperkenalkan untuk menyediakan medan penerimaan yang lebih pelbagai untuk menangkap wajah dalam pose dan oklusi yang ekstrem dengan lebih baik. Tambahan pula, rangkaian saraf konvolusional (CNN) ResNet digunakan sebagai tulang belakang model kami demi keupayaan pengesanan wajah yang lebih baik. Pendekatan pengesanan wajah terkini telah dikaji supaya lebih memahami sintesis pemikiran semasa dalam penyelidikan ini. Eksperimen ekstensif yang dilakukan pada set data penanda wajah WIDER FACE yang popular menunjukkan model kami mencapai hasil yang kompetitif dalam pengesanan wajah. RebifFace dinilai dengan purata ketepatan (AP) dan mencapai 86.6% (mudah), 84.2% (sederhana), dan 66.5% (sukar) dalam set pengesahan.

1 PENGENALAN

Rajah 1.1 Himpunan Orang Ramai yang Padat. Masalah variasi skala (Kotak Hijau) dan oklusi (Kotak Merah) dalam gambar himpunan orang ramai

Sumber: https://www.huffingtonpost.ca/2014/05/16/canadian-summer-music-festivals_n_5144202.html

Pengesanan wajah kecil dalam himpunan orang ramai yang padat merupakan salah satu masalah yang mencabar dalam domain penglihatan komputer. Tujuan pengesanan wajah adalah untuk mengenal pasti wajah dalam gambar dan mengembalikan lokasi wajah dalam gambar. Dalam himpunan orang ramai yang padat seperti yang ditunjukkan dalam Rajah 1.1, bilangan individu yang banyak menyebabkan oklusi teruk antara individu dan skala wajah yang berbeza, membawa masalah maklumat semantik konteks (terdiri daripada lokasi, tekstur, dan skala spasial) yang tidak mencukupi dalam ciri-ciri yang diekstrak untuk pengesanan wajah yang tepat.

Dengan perkembangan pesat dalam rangkaian saraf konvolusional (CNN) yang mendalam, pengesanan wajah baru-baru ini telah mencapai kemajuan yang menakjubkan. Prestasi purata ketepatan pada set data WIDER FACE yang mencabar telah meningkat dari 40% kepada 90% sejak beberapa tahun kebelakangan ini. Beberapa karya seperti Aggregate Channel Features (Bin Yang et al. 2014) dan A Deep Leaning Approach (Shuo Yang et al. 2015) melaksanakan kaedah pembelajaran mendalam dengan kombinasi kaedah tradisional (contohnya, deformable part models, cascade structure) untuk melakukan pengesanan wajah. Dalam Mask R-CNN (K. He et al. 2017), anotasi piksel yang padat meningkatkan prestasi pengesanan dengan ketara. Malangnya, anotasi wajah yang padat tidak mungkin dilakukan dalam set data wajah yang mencabar seperti WIDER FACE . (Shuo Yang et al. 2016)

Dalam projek ini, satu pendekatan berdasarkan CNN akan dicadangkan untuk mencapai prestasi tinggi dalam pengesanan wajah dalam himpunan orang ramai. Dalam model yang dicadangkan, ciri-ciri yang penting untuk mengesan wajah yang pelbagai skala akan diekstrak. Model yang dicadangkan akan dilatih dengan strategi pembelajaran pelbagai tugas untuk mengekalkan ketepatannya. Prestasi pengesanan wajah akan dinilai dengan purata ketepatan (AP). Pendekatan kami akan dinilai dengan menggunakan set data wajah tanda aras awam WIDER FACE.

2 PENYATAAN MASALAH

Selama bertahun-tahun, penyelidik penglihatan komputer telah meningkatkan prestasi algoritma pengesanan wajah dalam pelbagai tugas pengesanan wajah yang mencabar. Prestasi purata ketepatan (AP) algoritma pengesanan wajah canggih pada set data WIDER FACE yang mencabar adalah sekitar 90% (sekitar 900 wajah daripada 1000 wajah dalam himpunan orang ramai akan dikesan) sejak beberapa tahun kebelakangan ini. Cabaran sebenar dalam teknologi pengesanan wajah adalah keupayaan untuk menangani semua senario dan keadaan di mana subjek tidak bekerjasama. Terdapat banyak faktor yang menyebabkan penampilan wajah berbeza-beza. Variasi skala dan oklusi seperti yang ditunjukkan dalam Rajah 1.1 dianggap sebagai salah satu cabaran paling kritikal dalam pengesanan wajah. Pengesan wajah sukar untuk berurusan dengan jangka skala yang besar. Tambahan pula, konteks maklumat semantik wajah kecil adalah tidak jelas terutamanya wajah oklusi. Oleh itu, tidak mudah untuk mengekstrak ciri yang penting daripada wajah kecil dalam pengesanan wajah. Pendekatan sedia ada menggunakan teknik pengekstrakan ciri buatan tangan atau CNN mendalam untuk mengekstrak ciri tahap rendah dan tinggi untuk latihan. Walaupun banyak kaedah semasa mencapai prestasi yang baik dalam penanda aras awam, tetapi mereka masih sukar mengekalkan ketepatan mereka dalam himpunan orang ramai. Oleh itu, projek ini dijalankan untuk mencadangkan penyelesaian untuk mengesan wajah dengan pelbagai skala dan oklusi dalam gambar himpunan orang ramai.

3 OBJEKTIF KAJIAN

Tujuan projek ini adalah untuk mencadangkan algoritma / model untuk mengesan wajah pada skala dan oklusi yang berlainan dalam himpunan orang ramai. Terdapat tiga objektif dalam projek ini untuk mencapai tujuan yang dinyatakan:

i. Mencadangkan satu seni bina rangkaian saraf konvolusional mendalam untuk mempelajari ciri-ciri penting wajah dan mengesan wajah dalam gambar himpunan orang ramai.

ii. Melokalisasi wajah yang dikesan dalam gambar himpunan orang ramai dengan kotak pengikat.

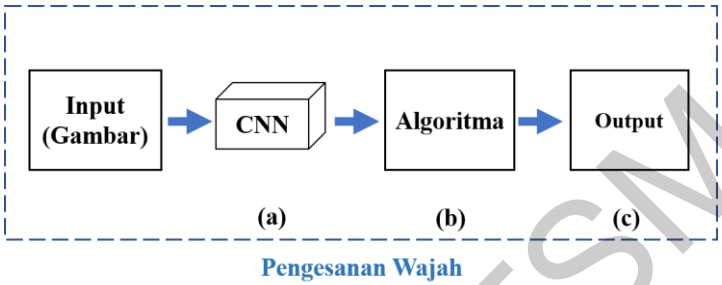
iii. Menguji model yang dicadangkan dengan set data wajah tanda aras awam untuk menentukan kebarangkalian dan ketepatan hasil yang diramalkan.

4 METOD KAJIAN

Pengesanan wajah dalam projek ini merupakan process mengesan dan melokalisasi wajah secara automatik dalam gambar himpunan orang ramai. Ini adalah tugas yang mencabar disebabkan masalah seperti variasi skala, oklusi, dan pencahayaan. Antara cabarannya, projek ini akan memberi tumpuan kepada masalah variasi skala dan oklusi. Oleh itu, model kami memerlukan seni bina dan kaedah yang kuat untuk melaksanakan tujuan tersebut. Terdapat dua keadah reka bentuk yang terkenal dalam pengesanan wajah: Satu Tahap dan Dua Tahap. Reka bentuk satu tahap terkenal dalam pendekatan yang terkini. Ia sampel lokasi dan skala wajah secara padat pada *feature pyramid* (FP), menunjukkan prestasi yang menjanjikan dan menghasilkan kelajuan yang lebih cepat dibandingkan dengan kaedah dua tahap. Mengikut laluan ini, kami meningkatkan kerangka pengesanan wajah satu peringkat dan mencadangkan model pengesanan wajah yang bernama RebifFace dengan mengeksploitasi kerugian pelbagai tugas yang datang dari isyarat tambahan yang diawasi. Dalam bahagian ini, seni bina RebifFace dan setiap komponenya akan dibincangkan mengikut fungsi dan tujuannya. Secara amnya, RebifFace menggunakan rangkaian saraf konvolusional (CNN) mendalam dalam reka bentuknya dan seni bina RebifFace adalah berdasarkan pada FP yang membolehkan model mengesan wajah di sebilangan besar skala. Tambahan pula, fungsi

kerugian yang digunakan untuk mengukur seberapa baik RebifFace melaksanakan tugas pengesanan wajah dalam projek ini juga akan dibincangkan dalam bahagian ini.

4.1 STRUKTUR REBIFFACE SECARA KESELURUHAN

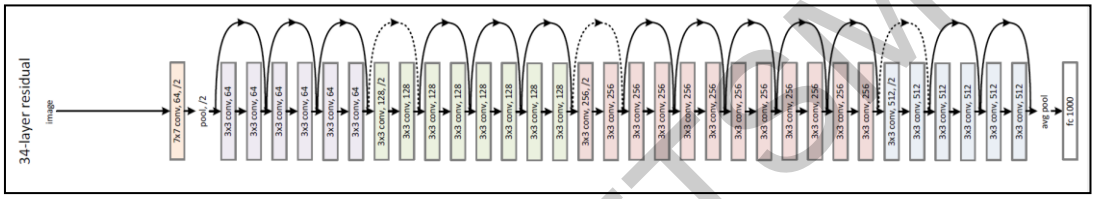


Rajah 4.1 Gambaran Keseluruhan Struktur RebifFace. (a) Tulang Belakang CNN. (b) Algoritma pengesanan wajah yang dicadangkan. (c) Output wajah yang dikesan dengan lokasinya.

Seperti yang ditunjukkan dalam Rajah 4.1, proses keseluruhan dalam struktur RebifFace terdiri daripada input gambar, tulang belakang CNN, algoritma pengesanan wajah kami dan hasil output. Input gambar adalah proses pertama yang memuatkan gambar ke dalam RebifFace. Gambar yang akan dimuatkan terdiri daripada gambar RGB (Merah, Hijau, Biru) yang mempunyai tiga saluran warna atau gambar skala kelabu yang mempunyai satu saluran warna. Seterusnya, gambar dimuat ke tulang belakang CNN untuk pengekstrakan ciri. CNN yang digunakan sebagai tulang belakang RebifFace ialah rangkaian klasifikasi pra-latihan ResNet-50 (Kaiming He et al., 2016) yang telah dilatih pada set data ImageNet-11k. ResNet telah mencapai hasil prestasi yang luar biasa dalam pendekatan yang terkini dalam tugas pengesanan wajah. Dengan memanfaatkan pemetaan sisa dalam ResNet, RebifFace dapat memiliki rangkaian saraf yang lebih mendalam dalam modelnya. Data gambar kemudian dihantar ke algoritma pengesanan wajah RebifFace. Di sini, ciri-ciri paling deskriptif dan menonjol untuk wajah dalam himpunan orang ramai diekstrak. Secara khususnya, kami menerapkan pembelajaran transfer dengan menggunakan model pra-latihan ResNet-50 di mana kami tidak perlu melatih keseluruhan rangkaian saraf mendalam tetapi hanya lapisan yang baru ditambahkan (algoritma pengesanan RebifFace) untuk pengesanan wajah. Kemudiannya, hasil terakhir yang menunjukkan ketepatan dan kebarangkalian hasil ramalan dan posisi spasial wajah (x_{min}, y_{min}, lebar, ketinggian) yang dikesan dihasilkan sebagai output dari RebifFace. Lokasi wajah ditunjukkan dengan kotak pengikat.

Intersection over union (IoU) yang digunakan untuk mengukur ketepatan pengesanan wajah dihitung melalui kotak pengikat yang diramalkan dan kotak pengikat *ground-truth* pada set data wajah tanda aras awam. Purata ketepatan (AP) dikira sebagai metrik penilaian bagi RebifFace untuk menilai prestasi pengesanan wajahnya.

4.1.1 TULANG BELAKANG RANGKAIAN SARAF KONVOLUSIONAL (CNN) – SENI BINA RESNET



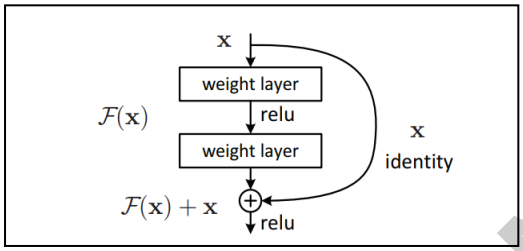
Rajah 4.2 Seni Bina Rangkaian Residual dengan 34 Lapisan Parameter (ResNet-34)

Sumber: https://arxiv.org/pdf/1512.03385.pdf

ResNet (Kaiming He et al., 2016) digunakan untuk meningkatkan kedalaman rangkaian saraf demi meningkatkan ketepatan model. Oleh itu, model yang dicadangkan dapat mengekstrak dan mempelajari lebih banyak ciri dan seterusnya meningkatkan ketepatannya. Tetapinya, peningkatan kedalaman akan menimbulkan masalah kecerunan lenyap/meletup dan degradasi. Kecerunan lenyap/meletup berlaku apabila derivatif atau kecerunan dalam rangkaian mendalam terlalu besar atau terlalu kecil. Sebaliknya, masalah degradasi adalah masalah ketepatan rangkaian menjadi tepu dan kemudian menurun dengan cepat dengan peningkatan kedalaman rangkaian. ResNet dapat berfungsi dengan baik dalam mengatasi masalah yang disebutkan.

Rajah 4.2 menunjukkan Rangkaian Residual (ResNet) dengan 34 lapisan parameter untuk melakukan konvolusi pada gambar input. Namun, lapisan ResNet boleh ditingkat sehingga 152 lapisan. ResNet kebanyakannya terdiri daripada penapis konvolusi 3x3 dan 7x7. Rangkaian ini diakhiri dengan lapisan purata pooling global dan lapisan bersambungan sepenuhnya 1000-arah dengan Softmax. ResNet menyelesaikan masalah yang disebutkan dengan menggunakan sambungan jalan pintas seperti yang ditunjukkan dalam Rajah 4.3 yang membenarkan model mempelajari fungsi mengenal pasti yang mengekalkan prestasi lapisan jaringan yang lebih tinggi sehebat dengan lapisan bawah sambil meningkatkan ketepatan keseluruhannya. Oleh itu, ResNet

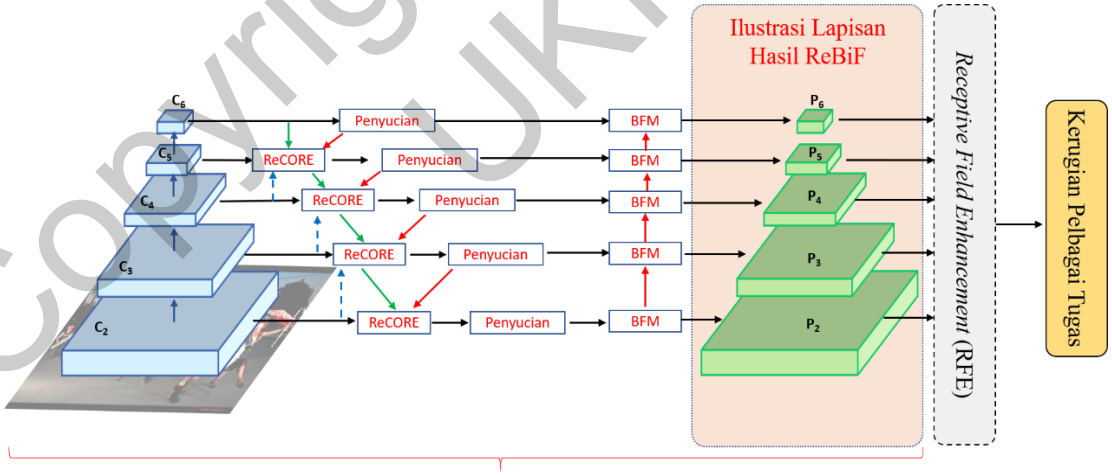
dengan 50 lapisan parameter digunakan sebagai rangkaian tulang belakang dalam projek ini. Rangkaian tulang belakang dimulakan oleh model pra-latihan ImageNet.



Rajah 4.3 Pembelajaran *Residual* : Block Bangunan

Rajah 4.3 menunjukkan blok bangunan ResNet. Blok bangunan ResNet dikenali sebagai blok *residual*. Teras blok *residual* adalah sambungan yang disebut "sambungan pintasan". Sambungan jalan pintas adalah sambungan yang melangkau satu atau lebih lapisan. Ia melakukan pemetaan identiti dan menambahkan output dari satu atau lebih dari satu lapisan sebelumnya ke output lapisan bertumpuk seperti yang ditunjukkan dalam Rajah 4.3.

4.2 REKA BENTUK SENI BINA REBIFFACE



Rajah 4.4 Reka bentuk seni bina RebifFace. RebifFace direka berdasarkan *Residual Bi-Fusion* (ReBiF) *feature pyramid* (FP) dengan modul *Receptive Field Enhancement* (RFE). Kerugian pelbagai tugas dikira untuk setiap sauh.

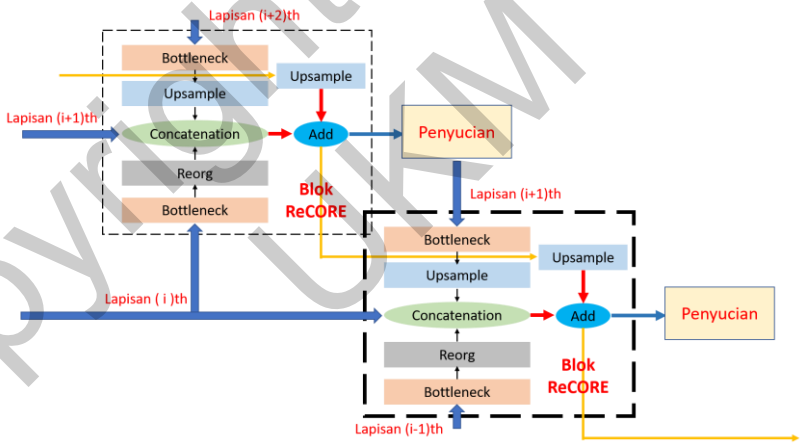
Seni bina RebifFace direka berdasarkan rangkaian Residual Bi-Fusion (ReBiF) feature pyramid (FP) (Ping-Yang Chen et al., 2019) dengan modul Receptive Field Enhancement (RFE) dan mengadopsi ResNet-50 sebagai tulang belakang seperti yang

ditunjukkan dalam Rajah 4.4. FP mengandungi peta ciri semantik peringkat tinggi pada semua skala yang membolehkan RebifFace mengekstrak ciri wajah yang penting dan mengesan wajah di sebilangan besar skala. Pelbagai FP yang canggih seperti Feature Pyramid Network (FPN) (Tsung Yi. Lin et al., 2017) menunjukkan hasil yang memberangsangkan pada pengesanan wajah bersaiz besar dan sederhana, tetapi menunjukkan prestasi yang sangat buruk pada wajah kecil. Hal ini kerana jalur dari atas ke bawah dalam FPN membawa kesan pergeseran pooling yang akan menyebabkan kegagalan dalam mengekalkan posisi wajah kecil yang tepat. Pengesanan wajah kecil amat mencabar dan memerlukan semantik tahap tinggi untuk membezakan wajah dari latar belakang dan ciri tahap rendah/pertengahan untuk penyetempatan wajah yang tepat. Untuk mengekalkan prestasi dalam pengesanan wajah bersaiz besar dan kecil, ReBiF FP digunakan dalam RebifFace. Kami meluaskan idea dari ReBiF asli tetapi dengan pertimbangan kami sendiri dalam projek ini. Ia adalah dua arah dan dapat menggabungkan kedua-dua ciri dalam dan cetek ke arah pengesanan wajah yang lebih berkesan dan kuat. Secara khususnya, ReBiF FP terdiri daripada modul ReCORE (Residual COncatenation and REorganization), modul Penyucian, dan modul Bottom-Up Fused (BFM). Dengan ReBiF, ia dapat mengimbangi maklumat yang hilang dari peta ciri lapisan bawah, menyimpan seberapa banyak ciri yang mungkin dan ekstrak ciri-ciri yang lebih penting untuk wajah dalam himpunan orang ramai.

Dalam RebifFace, ResNet diterap dengan struktur FP lima tingkat. RebifFace menggunakan tahap FP dari P2 hingga P6, di mana P2 hingga P5 dihitung dari blok residual ResNet yang dilambangkan sebagai C2, C3, C4, dan C5 yang masing-masing mempunyai ukuran ruang yang sama dengan menggunakan modul ReCORE, Penyucian dan BFM. C2 hingga C5 adalah rangkaian klasifikasi ResNet pra-latihan pada set data ImageNet-11k. P6 dihitung dari C6 yang diekstrak oleh lapisan konvolusi 3 x 3 sampel bawah dengan stride = 2 pada C5. Berlainan dengan P2 hingga P5, P6 dihitung melalui modul Penyucian dan BFM sahaja tanpa modul ReCORE untuk mengelakkan kerja duplikasi disebabkan oleh atribut C6 (ekstrak pada C5). Lapisan konvolusi yang baru ditambah, C6 dan P6 dimulakan secara rawak dengan kaedah “Xavier” (X. Glorot and Y. Bengio, 2010).

Sementara itu, RebifFace juga menerapkan modul RFE pada lima tingkat FP untuk meningkatkan medan penerimaan RebifFace dengan nisbah yang berbeza dan meningkatkan permodelan konteks yang tegar. Ini membantu dalam mengesan wajah dengan nisbah aspek dan tahap oklusi yang berbeza. Di samping itu, RebifFace menggunakan strategi pembelajaran pelbagai tugas untuk meramalkan skor wajah, kotak wajah, dan lima mercu tanda wajah secara serentak dalam tugas pengesanan wajah. Strategi pembelajaran pelbagai tugas terdiri daripada pembelajaran yang diawasi untuk meramalkan klasifikasi wajah dan regresi, dan pembelajaran pengawasan tambahan untuk meramalkan lima mercu tanda wajah pada wajah. RebifFace menghitung kerugian pelbagai tugas dalam strategi pembelajaran pelbagai tugas selepas modul RFE. Dengan kerugian pelbagai tugas, RebifFace dapat meramalkan wajah dalam gambar himpunan orang ramai dan mengeluarkan lokasi wajah yang dikesan dalam gambar.

4.2.1 MODUL ReCORE (*Residual Concatenation and Reorganization*)

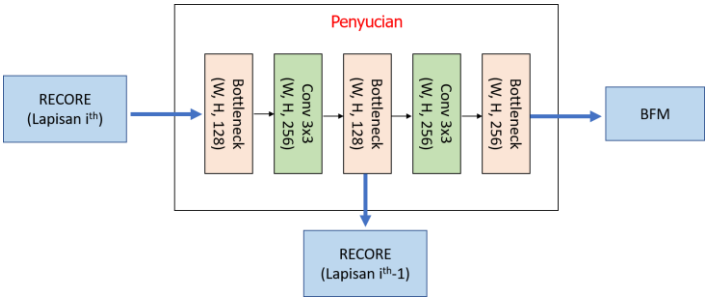


Rajah 4.5 Struktur modul ReCORE

Modul ReCORE digunakan pada setiap tahap ReBiF FP kecuali tahap terdalam di ReBiF FP seperti yang ditunjukkan dalam Rajah 4.4. Hal ini kerana tahap terdalam diekstrak dari tahap C_5 dalam projek ini. Sekiranya blok ReCORE digunakan pada tahap itu, ciri yang sama akan diduplikasi disebabkan pencampuran ciri dari C_5 dalam tahap sebelumnya. Oleh itu, sambungan lateral sahaja yang digunakan pada tahap terdalam dan sambungan lateral ini akan dihantar ke blok ReCORE terdalam dalam tahap (ith_{max} -1) sebelumnya. Blok ReCORE dihitung dengan menggunakan pendekatan atas-bawah di mana blok ReCORE tahap tertinggi akan dihitung terlebih dahulu sehingga blok

ReCORE tahap terendah. Seperti yang ditunjukkan dalam Rajah 4.5, setiap blok ReCORE menggabungkan pelbagai ciri dari tiga lapisan bersebelahan (sebelumnya, semasa, dan berikutnya) tulang belakang untuk memperkaya ciri wajah. Modul ReCORE dapat dilaksanakan secara rekursif bukan hanya untuk menggabungkan ciri semantik tahap tinggi dari lapisan dalam ke lapisan cetek (arah atas-bawah) tetapi juga menyusun semula ciri-ciri yang lebih kaya secara spasial dari lapisan cetik ke lapisan yang lebih dalam (arah bawah-atas). Modul ReCORE menggunakan operasi *concatenation* untuk menggabungkan ciri lapisan yang lebih dalam dan lapisan yang lebih cetek ke lapisan semasa. Untuk menjalankan operasi *concatenation*, lapisan yang lebih dalam disampel oleh lapisan konvolusi 3×3 sementara lapisan bawah diubah saiznnya kepada ukuran lapisan semasa dengan lapisan konvolusi 3×3. Seterusnya, output blok ReCORE pada setiap tahap FP disucikan oleh modul Penyucian. Output hasil dari modul (*i*+1)th ReCORE dan modul Penyucian dimasukkan ke modul ReCORE *i*th secara berulang untuk menghasilkan lebih banyak maklumat konteks semantik. Diilhamkan oleh konsep ResNet, formulasi rekursif "*residual*" diterapkan dalam modul ReCORE. Output (*i*+1) blok ReCORE disuntik secara rekursif ke blok ReCORE *i*th secara *element-wise*. Dengan konsep ini, ReBiF FP dapat mengedarkan maklumat wajah semantik dan penyetempatan secara dua arah dari kedua-dua lapisan dalam dan cetek. Dengan ini, ketepatan RebifFace dapat ditingkatkan.

4.2.2 MODUL PENYUCIAN

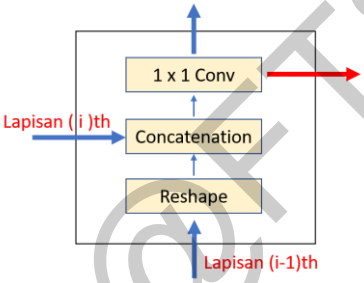


Rajah 4.6 Struktur modul Penyucian

Modul Penyucian digunakan untuk menyucikan output modul ReCORE untuk membentuk lebih banyak ciri-ciri kontekstual dan semantik. Rajah 4.6 menggambarkan saluran paip modul Penyucian. Modul ini terdiri daripada dua bahagian pengekstrakan ciri berturut-turut dan satu lapisan *bottleneck* dengan 256 saluran. Bahagian pengekstrakan ciri merangkumi satu lapisan *bottleneck* dengan 128 saluran dan lapisan

konvolusi 3×3. Dalam bahagian pengekstrakan ciri ini, lapisan *bottleneck* digunakan untuk mengurangkan bilangan saluran ke 128 saluran manakala lapisan konvolusi 3×3 digunakan untuk mengekstrak ciri-ciri kontekstual. Output lapisan *bottleneck* yang kedua dalam modul ini disuapkan ke modul ReCORE lain untuk menyempurnakan maklumat penyetempatan dan mencampurkan ciri-ciri pada skala yang lebih cetek. Akhir sekali, output lapisan *bottleneck* terakhir dalam modul Penyucian dimasukkan ke modul BFM.

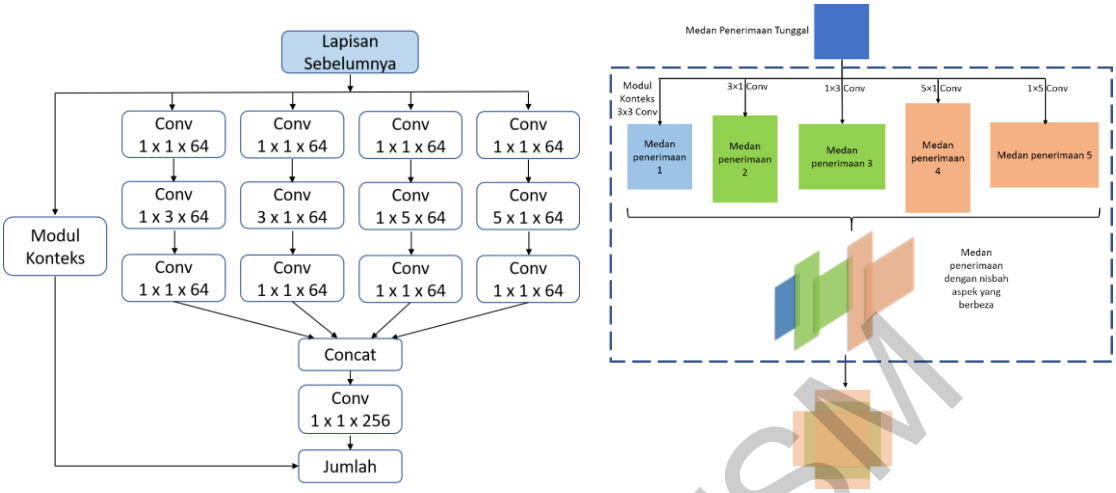
4.2.3 MODUL *BOTTOM-UP FUSED* (BFM)



Rajah 4.7 Struktur modul BFM

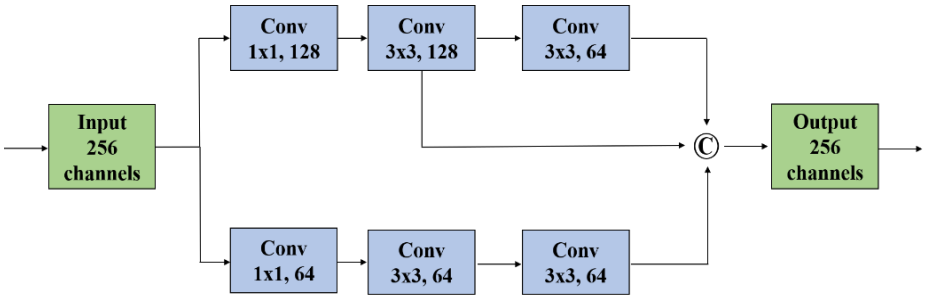
Modul ReCORE dan Penyucian dilaksanakan untuk meningkatkan ketepatan pengesanan wajah kecil. Namun begitu, peningkatan ini mungkin mengurangkan ketepatan pengesanan wajah bersaiz besar. Untuk mencapai kedua-dua ketepatan tinggi pada pengesanan wajah bersaiz kecil dan besar, ciri-ciri di antara lapisan dalam FP digabungkan lebit lanjut menggunakan jalur bawah ke atas dengan BFM. Seperti yang ditunjukkan dalam Rajah 4.7, lapisan semasa digabungkan dengan lapisan sebelumnya. Lapisan sebelumnya dibentuk semula oleh lapisan konvolusi 3×3 sample bawah agar sepadan dengan ukuran lapisan semasa sebelum penggabungan. Selepas itu, lapisan konvolusi 1×1 diterapkan pada output untuk pengurangan dimensi yang mengurangkan saluran ke 256 saluran. Hasil dari lapisan konvolusi 1×1 ini akan menjadi peta ciri output pada tahap semasa ReBiF FP. Hasilnya juga dimasukkan ke modul BFM lain untuk pencampuran ciri pada skala yang lebih dalam. Terdapat satu pengecualian dalam modul BFM, di mana lapisan paling rendah/cetek dalam ReBiF FP hanya akan melalui lapisan konvolusi 1×1 kerana tidak ada lapisan sebelum dalam tahap ini.

4.2.4 MODUL *RECEPTIVE FIELD ENHANCEMENT* (RFE)



Rajah 4.8 Struktur modul dan Ilustrasi RFE

Kebanyakan pendekatan pengesanan wajah sedia ada mempunyai medan penerimaan segi empat sama dalam model masing-masing. Modul pengekstrakan ciri asas dengan medan penerimaan segi empat sama boleh berfungsi dengan baik pada wajah yang bernisbah aspek kira-kira 1:1, tetapi tidak sesuai untuk model yang akan melatih set data yang mengandungi wajah pelbagai skala. Sekiranya terdapat ketidakcocokan antara medan penerimaan dan nisbah aspek wajah, ia akan menyebabkan model gagal mengesan wajah yang berpotensi. Diilhamkan oleh RefineFace (Shifeng Zhang et al. 2020), projek kami menggunakan modul bernama Receptive Field Enhancement (RFE) untuk menghadapi masalah tersebut. Modul RFE ini menggunakan struktur empat cabang seperti yang ditunjukkan dalam Rajah 4.8. Secara terperinci, lapisan konvolusi 1×1 diterapkan pada lapisan sebelumnya untuk mengurangkan ukuran saluran kepada 64 saluran. Seterusnya, lapisan konvolusi 1×k dan k×1 (di mana k = 3 dan 5) digunakan untuk menyediakan medan penerimaan segi empat tepat. Melalui lapisan konvolusi 1×1 yang lain, peta ciri dari empat cabang digabungkan bersama.



Rajah 4.9 Struktur Modul Konteks

Berlainan dengan RFE asli dalam RefineFace, modul konteks diterapkan secara langsung pada lapisan sebelumnya sebagai jalan tambahan di RFE kami. Reka bentuk modul Konteks digambarkan dalam Rajah 4.9. Modul konteks diterapkan untuk meningkatkan medan penerimaan segi empat sama dan meningkatkan daya pemodelan konteks yang tegar. Terdapat dua cabang dalam modul konteks. Cabang pertama menggunakan konvolusi 1×1 untuk mengurangkan ukuran saluran input ke 128 saluran dan dua lapisan konvolusi 3×3 seterusnya. Cabang kedua menggunakan konvolusi 1×1 untuk mengurangkan saiz saluran input ke 64 saluran dan dua lapisan konvolusi 3×3 seterusnya. Output dari kedua-dua cabangan akan digabungkan bersama dan menjadi saluran = 256. Output modul konteks ini seterusnya dijumlahkan dengan empat cabang lain di RFE. Seperti yang digambarkan dalam Rajah 4.8, modul RFE bersama dengan modul konteks menyediakan pelbagai medan penerimaan yang berguna untuk mengesan wajah berpose ekstrem.

4.2.5 KERUGIAN PELBAGAI TUGAS

Kerugian pelbagai tugas dalam pengesanan wajah RebifFace berasal dari kerangka kerja kerugian pelbagai dalam RetinaFace (JianKang Deng et al., 2020). Untuk sebarangan latihan sauh *i*, kerugian pelbagai tugas, *L* ditarifkan sebagai:

$$L = L_{cls}(p_i, p_i^*) + \lambda_1 p_i^* L_{box}(t_i, t_i^*) + \lambda_2 p_i^* L_{pts}(l_i, l_i^*) \quad \text{...(4.1)}$$

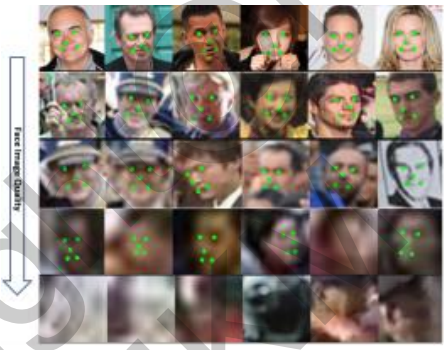
$$L_{cls}(p_i, p_i^*) = -\sum p_i^* \log(p_i) \quad \text{...(4.2)}$$

$$R(x) = smooth_{L1}(x) = \begin{cases} 0.5x^2, & if\ |x| < 1 \\ |x| - 0.5, & otherwise. \end{cases} \quad \text{...(4.3)}$$

(1) Kerugian klasifikasi wajah $L_{cls}(p_i, p_i^*)$, di mana p_i adalah ramalan kebarangkalian sauh *i* menjadi wajah dan p_i^* adalah *ground-truth* sauh *i*. p_i^* adalah 0 untuk sauh negatif dan 1 untuk sauh positif. Kerugian klasifikasi L_{cls} adalah kerugian *Softmax* untuk kelas binari (wajar / bukan wajah) seperti yang ditunjukkan dalam Persamaan 4.2. (2) Kerugian regresi kotak wajah $L_{box}(t_i , t_i^*)$, di mana $t_i = \{ t_x , t_y , t_w , t_h\}_i$ dan $t_i^* = \{ t_x^* , t_y^* , t_w^* , t_h^*\}_i$ mewakili koordinat kotak pengikat yang diramalkan dan kotak *ground-truth* berkaitan dengan sauh positive. Kami mengikut kaedah parameterisasi dalam *Fast*

RCNN (Ross Girshick, 2015) untuk menormalkan sasaran regresi kotak dan menggunakan $L_{box}(t_i , t_i^*) = R(t_i - t_i^*)$, di mana R adalah fungsi kerugian mantap (*smooth-L_1*) yang ditarifkan dalam *Fast-RCNN* seperti yang ditunjukkan dalam Persamaan 4.3. (3) Kerugian regresi mercu tanda wajah L_{pts} (l_i , l_i^*), di mana $l_i = \{l_{x1} , l_{y1} , \dots , l_{x5} , l_{y5}\}_i$ dan $l_i^* = \{l_{x1}^* , l_{y1}^* , \dots , l_{x5}^* , l_{y5}^*\}_i$. l_i mewakili lima mercu tanda yang diramalkan sementara l_i^* mewakili *ground-truth* yang berkaitan dengan sauh positif. Serupa dengan regresi kotak, regresi lima mercu tanda wajah juga menggunakan normalisasi sasaran berdasarkan pusat sauh. Parameter pengimbang kerugian λ_1 dan λ_2 ditetapkan ke 0.25 dan 0.1. λ_1 dan λ_2 digunakan untuk meningkatkan kepentingan kotak yang lebih baik dan lokasi mercu tanda.

4.3 PENGAWASAN TAMBAHAN



Rajah 4.10 Anotasi lima mercu tanda wajah pada wajah yang mampu dianotasikan daripada set latihan dan pengesahan WIDER FACE

Sumber: https://arxiv.org/pdf/1905.00641.pdf

Dalam projek ini, RebifFace mengesan wajah dan lima tanda wajah secara serentak. Pengawasan tambahan terhadap lima mercu tanda wajah ini akan meningkatkan ketepatan pengesanan wajah kecil. Lima mercu tanda wajah terletak di pusat mata (2 mercu tanda), hujung hidung (1 mercu tanda) dan sudut mulut (2 mercu tanda) seperti yang ditunjukkan dalam Rajah 4.10. Anotasi tambahan mengenai lima mercu tanda wajah pada gambar wajah daripada set latihan dan pengesahan dalam *WIDER FACE* yang disediakan dalam *RetinaFace* akan digunakan dalam projek ini. RebifFace akan meramalkan lima mercu tanda wajah dan membandingkan dengan mercu tanda *ground-truth* untuk mengira kerugian regresi mercu tanda wajah, L_{pts}.

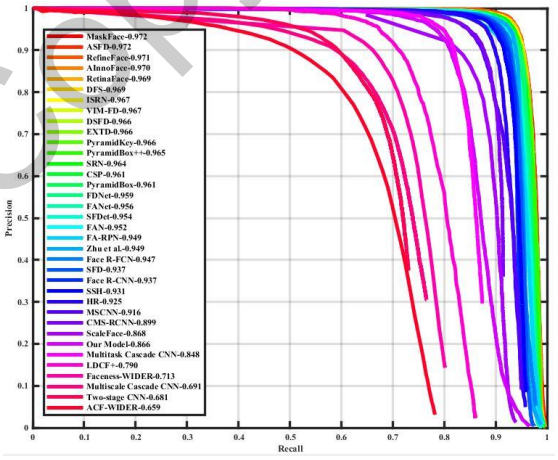
4.4 PENGOPTIMUM

Pengoptimum *stochastic gradient descent* (SGD) digunakan untuk melatih RebifFace dengan 0.9 momentum, 0.0005 weight decay dan 4 batch size pada set data *WIDER FACE*. Kami melatih RebifFace dengan satu NVIDIA Tesla T4 GPU (16GB) di Google Colab. Kami menggunakan strategi pemanasan untuk meningkatkan kadar pembelajaran secara beransur-ansur dari 10^{-3}, meningkat ke 10^{-2} selepas 5 epochs, kemudian dibahagi dengan 10 pada 55 dan 68 epochs.

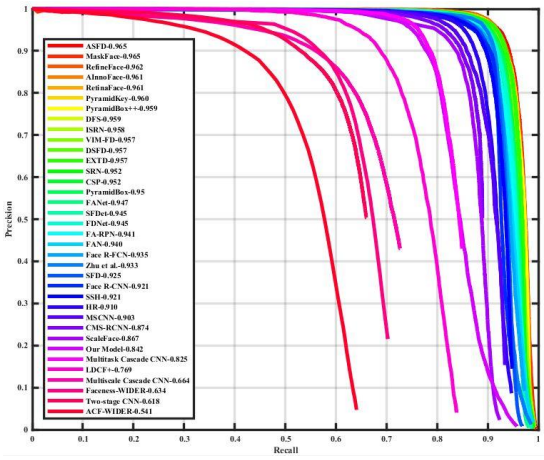
5 HASIL KAJIAN

5.1 HASIL PENILAIAN PENGESANAN WAJAH DALAM WIDER FACE

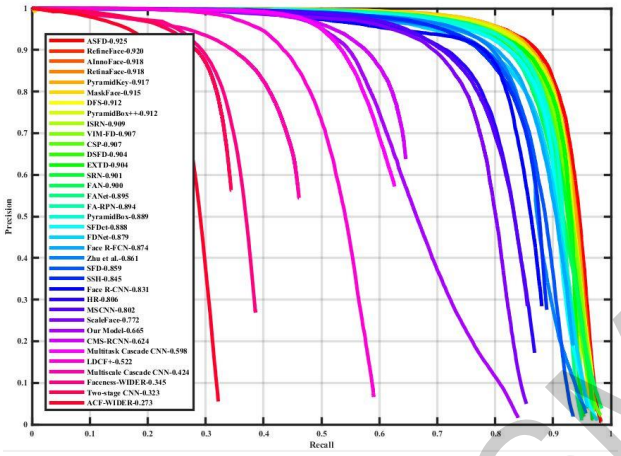
Untuk menilai prestasi RebifFace, kami membandingkan RebifFace dengan kaedah canggih termasuk RefineFace, RetinaFace, ScaleFace (S. Yang et al. 2017), dan Multitask Cascade (K Zhang et al. 2016). Rajah 5.1 menunjukkan perbandingan RebifFace dengan pendekatan canggih yang lain pada set data pengesahan berdasarkan lengkung *precision-recall* dan skor purata ketepatan (AP). Setiap subset WIDER FACE dibahagikan kepada tiga tahap kesukaran: Mudah, Sederhana, dan Sukar. Oleh itu, RebifFace diuji berdasarkan tahap kesukaran tersebut. RebifFace mencapai 86.6% (mudah), 84.2 % (sederhana), dan 66.5% (sukar).



(a) Val: Mudah (*Our Model*)



(b) Val: Sederhana (*Our Model*)



(c) Val: Sukar (*Our Model*)

Rajah 5.1 Lengkung *Precision-recall* pada set data pengesahan WIDER FACE

Jadual 5.1 Perbandingan Purata Ketepatan (AP) di antara RebifFace dan pendekatan canggih pada set pengesahan WIDER FACE

Model	Mudah	Sederhana	Sukar
RefineFace	97.100	96.200	92.000
RetinaFace	96.900	96.100	91.800
ScaleFace	86.800	86.700	77.200
RebifFace (Modul kami)	**86.600**	**84.200**	**66.500**
Multitask Cascade	84.800	82.500	59.800

Menurut Rajah 5.1, RebifFace mencapai prestasi yang menjanjikan pada set pengesahan WIDER FACE. Prestasi RebifFace sangat dekat dengan ScaleFace dan Multitask Cascade seperti yang ditunjukkan dalam Jadual 5.1. ScaleFace adalah pendekatan yang menerapkan pengesan variasi skala untuk menangani invari skala manakala Multitask Cascade adalah pendekatan yang menerapkan *feature pyramid*, mercu tanda wajah, *Proposal Network* (P-Net), dan *Refine Network* (R-Net) (K Zhang et al. 2016). Pendekatan tersebut berlainanan dengan RebifFace yang menerapkan ReBiF FP dan RFE dalam tugas pengesanan wajah. Walau bagaimanapun, RebifFace menunjukkan hasil yang kompetitif apabila dibandingkan dengan beberapa pendekatan yang canggih. Hasil ini menyakinkan kami bahawa modul ReBiF FP dan RFE berkesan dalam mengekstrak ciri-ciri penting wajah dengan masalah variasi skala dan oklusi. Sebaliknya, RebifFace berprestasi rendah berbanding dengan RefineFace dan RetinaFace. Ini mungkin disebabkan RefineFace dan RetinaFace menggunakan CNN

yang lebih mendalam, iaitu ResNet-152, sebagai tulang belakang model mereka. ResNet-152 mempunyai prestasi yang lebih baik berbanding dengan ResNet-50, yang digunakan dalam RebifFace. Lebih-lebih lagi, RefineFace dan RetinaFace telah dilatih selama 130 *epochs* dan 80 *epochs* masing-masing, yang lebih daripada 6 *epochs* dalam RebifFace.

5.2 ANALISIS MODEL

Untuk menganalisi keberkesanan RebifFace secara lanjut, kami membandingkan prestasi RebifFace dengan RetinaFace dengan *IoU threshold* 0.5 pada gambar yang sama. Kami memilih RetinaFace kerana ia mempunyai seni bina yang serupa (tetapi dengan teknik yang berbeza) dengan RebifFace, iaitu terdiri dari tulang belakang CNN ResNet, *feature pyramid*, dan pembelajaran pelbagai tugas mercu tanda wajah yang diawasi secara ekstra.

Rajah 5.2 Perbandingan dengan gambar selfie terbesar di dunia (1151 orang/wajah). (a) RetinaFace (~900 wajah) (b) RebifFace (155 wajah), (*IoU threshold* 0.5) daripada 1151 orang yang dilaporkan.

Semasa dibandingkan dengan gambar selfie terbesar di dunia dalam Rajah 5.2, RetinaFace dapat mengesan sekitar 900 wajah daripada 1151 orang yang dilaporkan sementara RebifFace hanya dapat mengesan 155 wajah. Hal ini menunjukkan RebifFace berprestasi rendah berbanding dengan RetinaFace dengan margin yang besar. RetinaFace menggunakan cabang penyahkod mesh melalui pembelajaran pengawasan diri yang dapat meramalkan bentuk wajah 3D piksel. Ia berfungsi dengan baik dalam meramalkan bentuk wajah 3D pada wajah yang sangat kecil atau oklusi. RebifFace gagal mengesan wajah yang terlalu kecil dan oklusi terutamanya wajah yang jauh dari kamera. Hal ini mungkin disebabkan oleh kekurangan latihan dalam RebifFace. RebifFace adalah model yang melatih sehingga 6 *epochs* sahaja selama 9

jam. Ia jauh lebih rendah daripada 80 *epochs* dalam RetinaFace. Kekurangan latihan menyebabkan RebifFace kurang mempelajari ciri wajah yang penting terutamanya wajah yang sangat kecil. Walau bagaimanapun, RebifFace masih menunjukkan hasil yang kompetitif. RebifFace dapat mengesan wajah yang dekat dengan pandangan kamera. Antara wajah yang dikesan, RebifFace berjaya mengesan wajah bersaiz besar, sederhana, dan kecil, wajah oklusi, dan wajah dengan aksesori fesyen seperti cermin mata hitam.

Rajah 5.3 Pengujian RebifFace secara manual pada gambar dalam set data ujian WIDER FACE

Selain itu, RebifFace juga diuji secara manual pada beberapa gambar yang mencabar dalam set data ujian WIDER FACE untuk mengenal pasti prestasi RebifFace. Kami membahagikan keputusan pengesanan wajah RebifFace kepada lima kategori iaitu, aksesori, kejelasan, pencahayaan, pose dan oklusi seperti yang ditunjukkan dalam Rajah 5.3 dan membandingkan kes pengesanan wajah yang berjaya dengan kes yang tidak berjaya. Dalam Rajah 5.3, kami perhatikan bahawa semua kategori ini mempunyai satu titik persamaan. Semua wajah yang mempunyai bentuk wajah dan ciri-ciri wajah yang jelas seperti mata, hidung dan mulut dapat dikesan manakala wajah yang tidak mempunyai bentuk wajah dan ciri-ciri wajah yang jelas amat sukar untuk dikesan. Hal ini menunjukkan bahawa RebifFace amat memerlukan ciri-ciri wajah yang jelas dalam pengesanan wajah.

Secara keseluruhannya, prestasi RebifFace selaras dengan objektif kami dalam projek ini. Hasilnya menunjukkan bahawa ReBiF FP berkesan dalam mengekalkan seberapa banyak ciri yang terdapat dalam peta ciri sementara mengekstrak ciri wajah yang penting dalam himpunan orang ramai. Dengan memanfaatkan pembelajaran

pelbagai tugas lima mercu tanda wajah yang diawasi secara ekstra dan kepelbagaian medan penerimaan, RebifFace melokalisasikan wajah yang dikesan dengan tepat.

6 KESIMPULAN

Kesimpulannya, projek ini bertujuan untuk menghasilkan satu model pembelajaran mendalam yang mampu mengesan wajah termasuk wajah kecil dalam gambar himpunan orang ramai. Kami mengkajikan cabaran bagi pengesanan wajah dalam gambar himpunan orang ramai dan mencadangkan penyelesaian satu peringkai, iaitu RebifFace, pada pengetahuan kami yang terbaik. Model yang dicadangkan dalam projek ini meningkatkan kemampuan regresi dan klasifikasi dalam pengesanan wajah untuk mengesan wajah pelbagai skala atau wajah oklusi. Pembangunan dan peningkatan ini telah memenuhi objektif kajian yang telah difokuskan dalam projek ini.

Namun begitu, terdapat beberapa kekangan dalam projek ini. Antaranya, kekangan yang paling kritikal adalah sumber pengkomputeran yang terhad terutamanya unit pemprosesan grafik (GPU). Ia menyebabkan RebifFace kurang dilatih dengan bilangan *epochs* yang cukup. Satu cadangan mengenai masalah ini telah dibincangkan demi penambahbaikan masa depan, iaitu melaksanakan projek ini pada unit GPU tempatan yang kuat dalam unit pengkomputeran. Dengan ini, kami dapat melatih RebifFace dengan bilangan *epochs* yang munasabah. Semua kekangan dan penambahbaikan masa depan akan dipertimbangkan untuk meningkatkan pretasi RebifFace.

7 RUJUKAN

Bin Yang, Junjie Yan, Zhen Lei, and Stan Z. Li, 2014. "*Aggregate Channel Features for Multi-view Face Detection*", IEEE International Joint Conference on Biometrics, pp. 1-8.

Chen, Ping-Yang, Jun-Wei Hsieh, Chien-Yao Wang, H. Liao and Munkhjargal Gochoo, 2019. "*Residual Bi-Fusion Feature Pyramid Network for Accurate Single-shot Object Detection.*" ArXiv abs/1911.12051.

J.K. Deng, J. Guo, Y. X. Zhou, J. K. Yu, I. Kotsia and S. Zafeiriou, 2020. "*RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild*", in CVPR.

Kaiming He, 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, pp. 770–778.

K. He, G. Gkioxari, P. Doll´ar, and R. Girshick, 2017. "*Mask R-CNN*", IEEE International Conference on Computer Vision (ICCV).

K Zhang, Z Zhang, Z Li, and Y Qiao, 2016. "*Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks*". arXiv preprint arXiv:1604.02878, 2016.

R. Girshick, 2015. "*Fast R-CNN*". In ICCV.

Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang, 2016. "*Wider face: A Face Detection Benchmark*". 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5525-5533.

Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang, 2015. "*From Facial Parts Responses to Face Detection: A Deep Learning Approach*", IEEE International Conference on Computer Vision (ICCV), pp.3676-3684.

S. Yang, Y. Xiong, C. C. Loy, and X. Tang, 2017. "*Face Detection through Scale-Friendly Deep Convolutional Networks*". arXiv preprint arXiv:1706.02863, 2017.

Tsung-Yi Lin, Piotr Doll´ar, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie, 2017. "*Feature Pyramid Networks for Object Detection*", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 936-944.

X. Glorot and Y. Bengio, 2010. "*Understanding the Difficulty of Training Deep Feedforward Neural Networks*". In AISTATS.

Zhang, Shifeng, Cheng Chi, Zhen Lei and S. Li, 2020. "*RefineFace: Refinement Neural Network for High Performance Face Detection*." IEEE transactions on pattern analysis and machine intelligence PP (2020)