BAB I

PENDAHULUAN

1.1 Latar Belakang

Manipulasi citra digital merupakan permasalahan yang sedang berkembang dalam masyarakat digital kita. Salah satu objek citra digital yang sering dimanipulasi adalah wajah. Dalam berkomunikasi wajah memainkan peran penting sebagai alat penunjang untuk menyampaikan informasi, karena wajah dapat menekankan pesan atau bahkan dapat menyampaikan pesan dengan sendirinya [1]. Wajah juga merupakan penanda dari identitas seseorang, di mana setiap wajah memiliki karakteristik khusus yang membedakan individu yang satu dengan yang lainnya. Dengan berkembangnya suatu ilmu teknologi pengolahan citra digital (*Digital Image Processing*) manipulasi citra wajah menjadi semakin realistis, sehingga membedakan citra digital telah dimanipulasi atau tidak menjadi sangat sulit dilakukan. Salah satunya adalah penggunaan *deepfakes* [1], yaitu sebuah teknik yang dapat mengubah wajah yang ada pada citra digital menjadi wajah orang lain dengan teknologi *Deep Learning*. Manipulasi citra wajah seperti ini sudah banyak diaplikasikan dalam kehidupan sehari-hari seperti pada aplikasi Instagram, Snapchat, Reface dan FaceApp untuk kebutuhan hiburan. Pada perkembangannya manipulasi citra wajah banyak dilakukan untuk tujuan kejahatan seperti untuk memfitnah seseorang dan pornografi.

Deepfakes dapat digunakan untuk menyebarkan berita palsu yang sulit untuk dibuktikan kepalsuannya sehingga akan memunculkan bentuk propaganda bentuk baru [2]. Konten digital yang dimanipulasi sangat mudah untuk menjangkau

masyarakat luas karena mudah disebarkan melalui media sosial dan situs *online*. Masyarakat yang menerima propaganda tersebut dapat dengan mudah terprovokasi sehingga dapat melakukan tindakan yang ekstrim baik secara ringan maupun berat. Teknologi seperti ini sangat berbahaya apabila digunakan untuk kepentingan politik karena dapat menjadi sumber kebingungan di tengah masyarakat. Pada konteks pornografi manipulasi citra wajah juga dapat dilakukan untuk memfitnah tokoh tertentu ataupun untuk melakukan *revenge porn*, yaitu tindakan mengganti wajah pemeran dalam video porno menjadi wajah orang lain. Dalam hal ini, menjadi penting untuk dapat mendeteksi citra wajah yang ada dalam video itu asli atau telah dipalsukan. Pada bidang *computer vision*, *machine learning* maupun *digital forensics* penelitian mengenai pendeteksian pemalsuan citra wajah masih terus dilakukan.

Beberapa penelitian yang berkaitan dengan deteksi pemalsuan citra wajah diantaranya menggunakan algoritma *Convolutional Neural Network* (CNN) dengan 6 arsitektur yang diajukan [1]. Arsitektur pertama bernama XceptionNet [3] dimana mempunyai 36 *convolution layer* dan disusun menjadi 14 modul yang semuanya memiliki koneksi residual linier di sekelilingnya, kecuali modul pertama dan terakhir. Arsitektur kedua sampai dengan ketujuh dengan nama Xcept. *Full Image* [1], Cozzolino *et al.* [4], Rahmouni *et al* [5], *Bayar and Stamm* [6] dan MesoNet [7], terdapat juga satu arsitektur yang tidak berbasis *Convolutional Neural Network* (CNN) yaitu Steg. Features + SVM. Penelitian ini membangun dataset pemalsuan wajah berskala besar baru bernama FaceForensics++ yang berisi lebih dari 1,8 juta gambar dari 4000 video palsu yang dibuat dengan algoritma pemalsuan citra wajah

DeepFakes [8], *FaceSwap* [9], *Face2Face* [10] dan *NeuralTexture* [11]. Pada penelitian tersebut akurasi tertinggi ditunjukan oleh arsitektur XceptionNet dengan total akurasi sebesar 70.10% pada dataset yang mereka kembangkan. Pada penelitian lainnya oleh [12], digunakan *Convolutional Neural Network* (CNN) untuk membandingkan jejak visual yang sangat kecil pada permukaan bingkai citra wajah asli dan palsu dan menyimpan fitur abnormal untuk pelatihan. Citra wajah dideteksi melalui arsitektur *Multi-task Cascaded Convolutional Networks* (MTCNN) dan lapisan *Convolutional Neural Network* (CNN) digunakan untuk mengekstrak fitur dan memberikan vector fitur ke sel LSTM sebagai input. Sebanyak 512 *landmark* wajah diekstraksi dan dibandingkan. Parameter seperti sinkronisasi bibir yang berkedip, pergerakan alis, dan posisi wajah adalah beberapa faktor penentu utama yang diklasifikasikan ke dalam data citra wajah asli atau palsu. Model di *training* pada jaringan Conv-LSTM yang menghasilkan *training accuracy* sebesar 95.12% dan *validation accuracy* sebesar 89.01%.

Berikutnya pada penelitian yang dilakukan oleh [13] membandingkan tingkat akurasi pada dua *dataset* yang berbeda yaitu FaceForensics++ dan Celeb-DF terhadap arsitektur XcepTemporal yang merka bangun. XcepTemporal merupakan pengembangan dari arsitektur XceptionNet dengan *Bidirectional* LSTM. Fitur dari XceptionNet dilewatkan ke lapisan *Bidirectional* LSTM pertama dan kedua untuk menghasilkan abstraksi fitur sekunder. Vektor fitur dari unit LSTM terakhir dari *bidirectional layer* ini diteruskan ke *fully-connected layer* dan akhirnya ke lapisan klasifikasi. *Dropout* ditambahkan ke *fully-connected layer* untuk *regularization*. XcepTemporal mempunyai 4 varian yaitu XcepTemporal

(CE), XcepTemporal (KL), XcepTemporal (EN) dan XcepTemporal (EN_{1+n}). Akurasi tertinggi berdasarkan *frame-level accuracy* pada dataset FaceForensics++ terdapat pada arsitektur XcepTemporal (KL) dengan akurasi 100% sedangkan pada dataset Celeb-DF terdapat pada arsitektur XcepTemporal (EN_{1+n}) dengan akurasi 97.83%. Komparasi dataset juga dilakukan pada penelitian [14], penelitian ini membandingkan akurasi pada dataset *Deepfakes Detection* (DFD), Celeb-DF dan *Deepfakes Detection Challenge* (DFDC) dengan menginisialisasi model XceptionNet dengan bobot pra-pelatihan ImageNet dengan metode *transfer learning* dan tanpa menggunakan *transfer learning*. *Transfer learning* bertujuan untuk memanfaatkan pengetahuan dari domain terkait (disebut domain sumber) untuk meningkatkan kinerja pembelajaran atau meminimalkan jumlah contoh berlabel yang diperlukan dalam domain target. Akurasi tertinggi dihasilkan pada dataset *Deepfakes Detection* (DFD) dengan metode *transfer learning* dimana akurasinya mencapai 95.86%, sedangkan dengan metode tanpa *transfer learning* mencapai akurasi sebesar 94.33%. Hasil penelitian menunjukan *transfer learning* dapat meningkatkan generalisasi pendeteksian sehingga meningkatkan akurasi. Berdasarkan penelitian oleh [15], akurasi yang dihasilkan oleh CNN dapat ditingkatkan dengan menggunakan metode *Greedy Layerwise Pretraining*. Penelitian tersebut menggunakan dataset berskala besar dari ImageNet dan CIFAR-10. Dalam perbandingan penelitian tersebut dengan penelitian lain didapatkan hasil bahwa hasil pengujian lebih baik dengan nilai akurasi sebesar 87.6%.

Pada penelitian yang dilakukan oleh [16], menggunakan arsitektur EfficientNet dan menemukan bahwa jumlah *frames per video* (FPV) dapat

mempengaruhi tingkat akurasi model. Saat menggunakan jumlah FPV yang sangat kecil, ada kecenderungan terjadinya penyesuain secara berlebihan pada arsitektur yang digunakan dan meningkatkan jumlah FPV tidak meningkatkan kinerja dengan cara yang dapat dibenarkan. Namun, memilih 32 FPV membantu mencegah *overfitting* pada penelitian yang diusulkan. Proses *training* dilakukan dengan menggunakan dataset FaceForensic++ dan *Deepfakes Detection Challenge* (DFDC) dan menghasilkan akurasi AUC terbaik sebesar 94.44% pada dataset FaceForensics++. Berikutnya penelitian mengenai deteksi citra wajah palsu dengan arsitektur Face X-ray [17] menemukan bahwa model yang dihasilkan *Convolutional Neural Network* (CNN) terutama arsitektur XceptionNet [3] menunjukan penurunan akurasi yang signifikan pada deteksi video *Deepfakes* yang dibuat dengan algoritma pemalsuan yang tidak ada di dalam dataset. Face X-ray diusulkan untuk mengatasi hal tersebut, arsitektur ini fokus pada *blending artifacts* untuk melakukan pendeteksi pemalsuan wajah yang lebih umum dan dapat dilatih tanpa citra palsu yang dibuat dengan salah satu algoritma manipulasi wajah yang canggih. Arsitektur ini pada dataset FaceForensics++ memberikan akurasi AUC sebesar 98.52% dan pada dataset *Deepfakes Detection* (DFD) sebesar 93.47%. Namun, arsitektur ini mempunyai limitasi dimana terlalu bergantung pada adanya langkah pencampuran (*blending step*) dan akurasi turun drastis pada citra wajah yang berkualitas rendah.

Berdasarkan uraian diatas, peneliti yang dilakukan oleh [1], [12]–[14], [16] melakukan penelitian deteksi manipulasi citra wajah dengan berbagai macam arsitektur *Convolutional Neural Network* (CNN), salah satunya adalah arsitektur

XceptionNet yang memberikan hasil akurasi tertinggi pada penelitian yang dilakukan. Namun, penelitian yang dilakukan oleh [17] menunjukan bahwa akurasi arsitektur *Convolutional Neural Network* (CNN) mengalami penurunan yang signifikan pada deteksi video *Deepfakes* yang dibuat dengan algoritma yang tidak ada di dalam dataset. Dari penelitian [1], [12]–[14], [16], [17] diketahui akurasi dipengaruhi oleh arsitektur, *dataset* dan *frame per video* (FPV) yang digunakan sehingga perlu penelitian yang lebih dalam mengenai hal tersebut. Oleh karena itu, penulis akan melakukan eksplorasi dan improvisasi arsitektur XceptionNet dalam pengenalan atau identifikasi citra wajah palsu menggunakan objek penelitian dari *dataset* FaceForensics++ [1], Celeb-DF [18] dan *Deepfakes Detection Challenge* (DFDC) [19]. Selain itu juga penulis akan melakukan eksplorasi jumlah *frame per video* (FPV) yang baik untuk digunakan. Penelitian ini diimplementasikan dengan menggunakan data uji video *deepfakes* dari YouTube.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah disampaikan, maka yang akan menjadi bahasan dalam penelitian ini adalah:

1. Bagaimana pengaruh arsitektur, *dataset* dan *frame per video* (FPV) terhadap tingkat akurasi pada arsitektur XceptionNet?
2. Bagaimana improvisasi arsitektur XceptionNet yang dapat dilakukan dalam pengenalan citra wajah yang dipalsukan dalam video?
3. Apakah akurasi yang dihasilkan oleh arsitektur XceptionNet untuk pengenalan citra wajah yang dipalsukan dalam video baik?

1.3 Batasan Masalah

Adapun batasan masalah yang digunakan agar penelitian ini dapat fokus pada pokok pembahasan dan tidak menyimpang dari tujuan penelitian. Berikut adalah batasan masalah yang ada dalam penelitian ini:

1. Data yang digunakan pada penelitian ini adalah data berupa video yang bersumber dari 3 *dataset* yang berbeda. Setiap *dataset* akan dibagi menjadi 3 bagian yaitu data *training*, *validation* dan *testing* sebanyak masing-masing 80%, 10% dan 10%. Selain itu dibuat juga *dataset* dari penggabungan 3 *dataset* untuk perbandingan dan generalisasi data. Penelitian ini tidak menggunakan keseluruhan *dataset* yang anda, digunakan metode *undersampling* pada *dataset* dengan detail sebagai berikut :
    a. FaceForensics++ [1] dengan memiliki total 5,000 video yang terdiri dari 1,000 video asli dan 4,000 video yang telah dimanipulasi dengan metode *DeepFakes* [8], *FaceSwap* [9], *Face2Face* [10] dan *NeuralTexture* [11]. Terdapat 3 kualitas video yang dipalsukan yaitu Raw, HQ dan LQ. Dalam penelitian ini akan menggunakan video berkualitas Raw terdiri dari 400 video asli dan 400 video yang telah dimanipulasi.
    b. Celeb-DF [18] terdiri dari 590 video asli dan 5,639 video *Deepfakes*. Video asli diambil dari YouTube yang terdiri dari video 59 selebriti dengan jenis kelamin, usia, etnis yang berbeda dan mempunyai kondisi pencahayaan yang berbeda. Dalam

penelitian ini akan menggunakan 400 video asli dan 400 video yang telah dimanipulasi.

c. *Deepfakes Detection Challenge* (DFDC) [19] memiliki total 5,000 video baik asli dan telah dimanipulasi yang terdiri dari 66 individu yang berbeda. Dalam penelitian ini akan menggunakan 400 video asli dan 400 video yang telah dimanipulasi.

d. *Combine*, merupakan *dataset* yang didapatkan melalui penggabungan ketiga *dataset* sebelumnya dengan hanya menggunakan 50% dari keseluruhan data untuk alasan komputasi. Sehingga memiliki total data 480 video asli dan 480 video yang telah dimanipulasi.

2. *Dataset* yang digunakan adalah video berformat .mp4 yang hanya berisi citra wajah berformat *Motion* JPEG (MJPEG) dengan resolusi 229 x 229 *pixel* dan berisi 30 *frame* citra wajah per detiknya.
3. Deteksi wajah menggunakan *Histogram of Oriented Gradients* (HOG) dari *library* dlib dan *Multi-task Cascaded Convolutional Networks* (MTCNN).
4. Pelatihan hanya menggunakan 2 kelas yaitu *real* (citra wajah yang asli) dan *fake* (citra wajah yang dipalsukan).
5. Model yang dibangun belum mempertimbangkan audio pada video *deepfakes*, sehingga model tidak dapat mendeteksi pemalsuan pada audio.

6. Pengujian sistem dilakukan dengan menggunakan 20 video *Deepfakes* yang diambil secara acak dari YouTube.

1.4 Maksud dan Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut:

1. Mengetahui pengaruh arsitektur, *dataset* dan *frame per video* (FPV) terhadap tingkat akurasi pada arsitektur XceptionNet yang dibangun.
2. Mengetahui improvisasi apa saja yang dapat dilakukan untuk dapat menghasilkan arsitektur XceptionNet yang baik dalam melakukan pengenalan citra wajah yang dipalsukan dalam video.
3. Mengetahui tingkat akurasi pengenalan citra wajah yang dipalsukan dalam video menggunakan arsitektur XceptionNet.

1.5 Manfaat Penelitian

Manfaat dari penelitian ini adalah sebagai berikut:

1. Mengetahui faktor yang dapat mempengaruhi tingkat akurasi dari arsitektur XceptionNet pada pengenalan citra wajah yang dipalsukan dalam video.
2. Mengetahui improvisasi arsitektur XceptionNet yang baik untuk pengenalan atau identifikasi citra wajah yang dipalsukan dalam video.
3. Dapat menjadi bahan acuan bagi peneliti yang akan melakukan penelitian di bidang *Deep Learning* dengan menggunakan arsitektur XceptionNet.
4. Mempermudah pengenalan atau identifikasi citra wajah yang dipalsukan dalam video.

1.6 Metode Penelitian

Metode yang digunakan dalam penelitian ini meliputi metode pendekatan penelitian, pemrosesan data, desain eksperimen, evaluasi dan pengujian sistem.

1.6.1 Pendekatan Penelitian

Pendekatan yang digunakan adalah pendekatan secara kuantitatif artinya informasi atau data yang disajikan adalah empiris, objektif, terukur, rasional dan sistematis berdasarkan hasil penelitian yang dilakukan. Penelitian bertujuan untuk memahami permasalahan secara lebih mendalam dan mengembangkan teori yang sudah ada yang didapatkan dari studi literatur dengan permasalahan yang ingin diselesaikan.

1.6.2 Studi Literatur

Studi literatur dilakukan dalam penelitian ini untuk dapat mengumpulkan data dan mempelajari informasi pendukung penelitian yang diperoleh dari buku, jurnal, skripsi, tutorial, dan berbagai informasi lain yang berkaitan dengan penelitian. Informasi yang berkaitan tersebut antara lain, proses pengolahan citra, *Deep Learning, Recognition, Machine Learning, Supervised Learning,* struktur dari arsitektur XceptionNet, serta *library* yang digunakan dalam membantu penelitian ini.

1.6.3 Identifikasi dan Perumusan Masalah

Tahapan penelitian di mana mengidentifikasi dan merumuskan masalah yang akan diteliti, dalam kasus ini perumusan masalah adalah membuat model yang dapat mengenali citra wajah yang dipalsukan dalam video menggunakan arsitektur XceptionNet.

1.6.4 Setting Penelitian

Pada tahap ini, dilakukan analisa dan penentuan fitur perangkat lunak eksperimen serta setting lingkungan penelitian. Fitur perangkat lunak eksperimen yang harus dipenuhi adalah bagaimana sistem mampu memproses video dan mampu mengidentifikasinya. Sedangkan setting lingkungan penelitian terdiri dari pengaturan perangkat keras dan perangkat lunak yang akan digunakan dalam eksperimen. Pada bagian perangkat keras meliputi spesifikasi komputer yang digunakan dalam penelitian, sedangkan pada bagian perangkat lunak meliputi bahasa pemrograman yang digunakan, sistem operasi maupun *software development environment* yang digunakan.

1.6.5 Pengumpulan dan Pemprosesan Data

Berdasarkan permasalahan yang anda, selanjutnya diambil dataset untuk dapat membuat model yang mampu mengenali citra wajah yang dipalsukan dalam sebuah video.

1.6.5.1 Pengumpulan Data

Berdasarkan studi literatur, didapatkan 4 *dataset* yang akan digunakan dalam penelitian ini yaitu *dataset* FaceForensics++ [1], Celeb-DF [18], *Deepfakes Detection Challenge* (DFDC) [19] dan *Combine*. *Dataset* tersebut terdiri dari 2 label atau kelas yaitu *real* (video dengan citra wajah yang asli) dan *fake* (video dengan citra wajah yang dipalsukan). *Dataset* dipisahkan menjadi 3 bagian yaitu *training, validation,* dan *testing*. Pada proses pengumpulan data pengujian, diambil video *deepfakes* sebanyak 20 video secara acak dari YouTube. Hal ini dimaksudkan

untuk dapat menguji akurasi yang didapatkan pada video *deepfakes* yang dibuat dengan metode yang tidak diketahui.

1.6.5.2 Pemprosesan Data

Tahap ini dilakukan perancangan mengenai bagaimana citra wajah dalam setiap *frame* video akan diekstraksi untuk selanjutnya masuk ke tahapan desain eksperimen. Sebelum dilakukan ekstraksi citra wajah dilakukan proses deteksi citra wajah terlebih dahulu dalam setiap *frame* video. Deteksi citra wajah dalam setiap *frame* video menggunakan dua metode, yaitu *Histogram of Oriented Gradients* (HOG) dari *library* dlib dan *Multi-task Cascaded Convolutional Networks* (MTCNN) terinspirasi dari penelitian yang dilakukan oleh [1], [12]. Kedua metode tersebut akan dibandingkan untuk dapat mencari metode deteksi wajah yang dapat menghasilkan akurasi terbaik.

1.6.6 Desain Eksperimen

Pada tahap metode desain eksperimen, dilakukan perancangan improvisasi arsitektur XceptionNet dengan menggunakan metode *transfer learning* dan penetapan parameter seperti jumlah *epoch*, jumlah *batch* serta *learning rate*. Beberapa arsitektur yang dibangun merupakan hasil improvisasi terhadap arsitektur XceptionNet dengan metode *Greedy Layerwise Pretraining* maupun XceptionNet sebagai *feature extraction*. Pada proses *training* akan dilakukan perbandingan penggunaan *dataset* yang diproses menggunakan metode deteksi wajah *Histogram of Oriented Gradients* (HOG) dari *library* dlib dan *Multi-task Cascaded Convolutional Networks* (MTCNN) untuk dapat melihat metode pemrosesan yang paling baik digunakan. Proses *training* juga akan menggunakan jumlah *frame per*

video (FPV) dan jenis *dataset* yang berbeda – beda untuk membuktikan pengaruhnya terhadap akurasi dan waktu *training* dari setiap arsitektur yang dibangun. Proses perancangan ini dilakukan berdasarkan hasil dari analisis studi literatur yang telah diperoleh.

1.6.7 Evaluasi

Evaluasi dilakukan untuk mengetahui hasil akurasi dan *overfitting* terhadap model yang didapatkan dari implementasi arsitektur XceptionNet yang dibuat. Evaluasi dilakukan untuk melihat faktor – faktor yang dapat mempengaruhi peningkatan akurasi dan mengurangi *overfitting*.

1.6.8 Metode Pengujian Sistem

Pada tahap metode pengujian sistem, dilakukan untuk memastikan program yang dibuat berjalan sesuai apa yang diharapkan, pengamatan *overfitting* dari arsitektur XceptionNet yang telah dibuat, pengujian terhadap model yang didapatkan untuk pengenalan citra wajah yang dimanipulasi dalam video menggunakan data pengujian video yang diambil acak dari YouTube.

1.7 Sistematika Penulisan

Pada bagian ini dituliskan urut-urutan dan sistematika penulisan yang dilakukan. Sistematika penulisan ini dijelaskan dengan ringkas sebagai berikut:

BAB I PENDAHULUAN

Pada bab ini berisi mengenai latar belakang, rumusan masalah, Batasan masalah, maksud dan tujuan penelitian, metode penelitian, dan sistematika penulisan.

BAB II LANDASAN TEORI

Bab ini berisi mengenai tinjauan pustaka yang menjadi rujukan serta memuat teori-teori yang dijadikan dasar dari penelitian ini tentang deteksi citra wajah yang dipalsukan. Tinjauan pustaka yang ada berkaitan dengan *Deep Learning, Recognition, Machine Learning, Supervised Learning* dan *Computer Vision* dalam klasifikasi citra menggunakan arsitektur *Convolutional Neural Network* (CNN). Teori – teori pendukung yang diperlukan untuk menyelesaikan masalah yang berhubungan dengan pengenalan atau identifikasi citra wajah yang dipalsukan menggunakan arsitektur XceptionNet.

BAB III METODOLOGI PENELITIAN

Pada bab ini berisikan tentang metode pengambilan *dataset*, pengambilan data uji dari YouTube, proses pemrosesan *dataset* menggunakan *Histogram of Oriented Gradients* (HOG) dari *library* dlib dan *Multi-task Cascaded Convolutional Networks* (MTCNN), proses *training* arsitektur XceptionNet, pengukuran tingkat akurasi terhadap model yang didapatkan dari implementasi arsitektur XceptionNet yang telah diimprovisasi, pengujian data uji terhadap model yang dibuat, pengamatan *overfitting* dan parameter – parameter arsitektur XceptionNet yang digunakan.

BAB IV HASIL DAN PEMBAHASAN

Pada bab ini berisikan tentang pembahasan implementasi dari arsitektur XceptionNet dengan menggunakan metode *transfer learning* dalam pengenalan citra wajah yang dipalsukan dalam video. Membahas pengamatan *overfitting* dan hal yang mempengaruhi tingkat akurasi model yang didapatkan dari arsitektur yang

dibuat, serta pengujian data uji berupa video *deepfakes* yang didapatkan dari YouTube.

BAB V KESIMPULAN

Bab ini berisikan tentang kesimpulan dan saran yang didapatkan dari hasil penelitian ini.

DAFTAR PUSTAKA

Bab ini berisikan tentang pustaka yang digunakan penulis sebagai acuan dan bahan dalam penelitian yang dilakukan.