BULETIN PAGELARAN MAHASISWA NASIONAL BIDANG TEKNOLOGI INFORMASI DAN KOMUNIKASI
ISSN: xxxxxx

Deteksi Kriminal pada Video Kamera Pengawas dengan *Multiple Instance Learning Vision Transformers*

M.Alif Al Hakim[1], Tengku Laras Malahayati[2], Rakha Abid Bangsawan[3], Laksmita Rahadianti[4]
[1]Fakultas Ilmu Komputer Universitas Indonesia, Depok, 16424, email: malif.al@ui.ac.id
[2]Fakultas Ilmu Komputer Universitas Indonesia, Depok, 16424, email: tengku.laras@ui.ac.id
[3]Fakultas Ilmu Komputer Universitas Indonesia, Depok, 16424, email: rakha.abid@ui.ac.id
[4]Fakultas Ilmu Komputer Universitas Indonesia, Depok, 16424, email: laksmita@cs.ui.ac.id
Corresponding Author: M.Alif Al Hakim

INTISARI — Kamera pengawas atau CCTV telah digunakan secara luas untuk memantau dan mencegah tindak kejahatan. Namun, efektivitasnya sering kali terbatas oleh faktor *human error*, di mana petugas pengawas keamanan mungkin tidak selalu fokus memantau area yang ditampilkan pada monitor CCTV. Kelalaian dalam pemantauan ini menciptakan celah dalam mendeteksi kejahatan dengan segera. Oleh karena itu, dibutuhkan sistem yang dapat membantu mendeteksi tindakan kriminal secara otomatis dan *real-time*. Penelitian ini berfokus pada peningkatan efisiensi dalam mendeteksi tindakan kriminal melalui pengawasan video dengan menggunakan *Video Vision Transformers* yang dilatih dengan teknik *Multiple Instance Learning* dan pendekatan *Sliding Window*. Penelitian ini mencoba mengatasi keterbatasan pengawasan CCTV manual yang rentan terhadap *human error* di Indonesia. Dataset UCF Crime yang berisi video tindakan kriminal singkat berdurasi sekitar 4 menit digunakan dalam penelitian ini. Selain itu, *dataset* juga ditambah dengan video pribadi yang didapatkan oleh tim peneliti. Model *Video Vision Transformers* digunakan untuk mengekstrak fitur dari dataset video, dan model *predictor* dilatih untuk memprediksi *anomaly score* atau tingkat kejanggalan terhadap aksi kriminal pada video. Sistem yang diusulkan mendeteksi aksi kriminal secara otomatis berdasarkan nilai *anomaly score* tiap *frame*. Metode yang diusulkan berhasil mencapai skor AUC (*Area Under the Curve*) sebesar 0.942 pada konteks *video level* dan 0.749 pada konteks *frame level*, menunjukkan efektivitas dalam mendeteksi tindakan kriminal dalam video kamera pengawas. Metode ini juga dibandingkan dengan metode ekstraksi fitur menggunakan model berbasis *Convolutional Networks* dengan dua *stream* gambar. Hasil penelitian ini diharapkan dapat memberikan kontribusi signifikan dalam meningkatkan kualitas pelayanan publik di bidang keamanan dengan mengurangi *human error*, meningkatkan efisiensi pengawasan, dan menciptakan lingkungan yang lebih aman dari tindak kejahatan.

KATA KUNCI — CCTV, Deteksi, Kriminal, *Multiple Instance Learning*, *Sliding Window*, *Video Vision Transformers*, *Weakly Supervised*

I. PENDAHULUAN

A. Latar Belakang Masalah

Rasa aman merupakan salah satu kebutuhan utama manusia. Rasa aman didefinisikan oleh Maslow, sebagaimana dikutip oleh Potter dan Perry, sebagai perasaan terlindungi dari ancaman atau teror dari luar dan dalam dirinya terkait dengan keamanan. Rasa aman merupakan kondisi di mana seseorang bebas dari cedera fisik dan psikologis dan dalam kondisi aman dan tenteram [1]. Rasa aman dapat diperoleh dari berbagai hal, salah satunya terbebas dari tindak kejahatan.

Tindak kejahatan di Indonesia merupakan peristiwa yang mengancam keamanan masyarakat dan melanggar hukum pidana. Terdapat beberapa contoh kasus kejahatan yang kerap terjadi di Indonesia, seperti pencurian, penculikan, pembegalan, penipuan hingga pembunuhan. Data kejahatan tahun 2024 (hingga minggu kedua bulan Mei 2024) yang diakses dari Kepolisian Negara Republik Indonesia memperlihatkan lima jenis tindak kejahatan yang meliputi pencurian dengan pemberatan sebanyak 18.447 kasus, pencurian biasa 15.267 kasus, penganiayaan 15.164 kasus, penipuan/perbuatan curang 13.931 kasus, dan kejahatan narkotika sebanyak 13.482 kasus [2].

Badan Pusat Statistik, dengan mengutip Kitab Undang-Undang Hukum Pidana (KUHP) Republik Indonesia dan International Classification of Crime for Statistical Purposes (ICCS) yang digagas oleh United Nations Office on Drug and Crime (UNODC), membuat klasifikasi tindak kejahatan sebagaimana yang ditampilkan pada Tabel 1.

Dilihat dari waktu terjadinya tindak kejahatan, rentang waktu antara pukul 18:00-21:59 merupakan saat di mana terjadinya tindak kejahatan tertinggi dengan 24.190 kasus. Tindak kejahatan juga banyak terjadi pada rentang waktu pukul 08:00-11:59 dengan jumlah kasus sebanyak 23.338 [2].

Situasi dan kondisi keamanan di Indonesia bergerak dinamis. Badan Pusat Statistik Indonesia dalam Publikasi Statistik Kriminal 2023 memperlihatkan bahwa pada tahun 2020 total jumlah kejadian kejahatan di Indonesia mencapai 247.218 kejadian. Angka tersebut mengalami penurunan pada tahun 2021 dengan 239.481 kejadian. Tren penurunan tersebut hanya berlangsung sesaat karena pada tahun 2022 terdapat 372.965 kejadian kejahatan. Lembaga tersebut juga menyatakan bahwa tingkat kejahatan (*crime rate*) menunjukkan pola yang sama.

Pada tahun 2020, tingkat kejahatan sebesar 94 dan pada tahun 2021 mengalami penurunan menjadi 91. Sayangnya, tingkat kejahatan tersebut kemudian meningkat menjadi 137 pada tahun 2022. Dalam Publikasi Statistik Kriminal 2023 juga dapat diperoleh informasi tentang interval waktu terjadinya kejahatan (*crime clock*), di mana pada tahun 2020 adalah 00.02'.07", lalu menjadi 00.02'.11" pada tahun 2021 dan

ISSN: xxxxxx

menjadi semakin pendek pada tahun 2022 menjadi 00.01'.24". Interval waktu terjadinya tindak kejahatan yang semakin memendek tersebut menunjukkan terjadinya peningkatan intensitas kejadian tindak kejahatan [3].

Relatif tingginya tingkat kejahatan di Indonesia sebagaimana diuraikan di atas menjadikan penjagaan keamanan dan pencegahan terjadinya tindak kejahatan menjadi penting, salah satunya dengan menggunakan perangkat kamera pengawas, atau yang biasa dikenal dengan Closed-Circuit Television (CCTV). Stutzer dan Zehnder, sebagaimana dikutip oleh Kurnia B.P. menyebutkan bahwa perangkat CCTV berperan sebagai pengawas dan pengintai yang dapat memantau situasi dan kondisi secara *real time* yang juga mampu menekan aksi kejahatan untuk terjadi [4]. Kurnia B.P. juga menyatakan bahwa CCTV digunakan pula oleh mereka yang bertugas sebagai petugas pengawas di mana mereka mengawasi CCTV dengan maksimal dalam rangka preventif untuk meminimalisir dan mendeteksi tindakan yang mencurigakan [4]. Perangkat CCTV memiliki peran yang signifikan dalam melawan dan mencegah terjadinya aksi tindak kejahatan serta memberikan perlindungan bagi masyarakat di ruang publik maupun privat. Penggunaan perangkat CCTV dianggap mampu memberikan jaminan keamanan bagi masyarakat sebagai bentuk upaya preventif dalam deteksi dini terjadinya aksi kejahatan dan situasi darurat.

Sayangnya, pengawasan dan pemantauan CCTV secara manual oleh petugas pengawas keamanan masih belum cukup efektif sebagai upaya deteksi dini, terutama untuk mendeteksi aksi kejahatan. Hal ini disebabkan oleh adanya faktor *human error* yang dapat terjadi ketika pengawasan dan pemantauan dilakukan secara manual. Sebagai contoh, petugas pengawas keamanan mungkin tidak selalu fokus memantau area yang ditampilkan pada monitor yang menampilkan rekaman video CCTV ketika sedang lelah. Kelalaian dalam memantau CCTV secara manual tersebut menciptakan celah dalam mendeteksi aksi kejahatan.

No.	Klasifikasi Kejahatan	Jenis Kejahatan
1.	Kejahatan terhadap nyawa	Pembunuhan
2.	Kejahatan terhadap fisik/badan	- Penganiayaan berat - Penganiayaan ringan - Kekerasan dalam rumah tangga
3.	Kejahatan terhadap kesusilaan	- Perkosaan - Pencabulan
4.	Kejahatan terhadap kemerdekaan orang	- Penculikan - Mempekerjakan anak di bawah umur
5.	Kejahatan terhadap hak milik/barang dengan penggunaan kekerasan	- Pencurian dengan kekerasan - Pencurian dengan kekerasan menggunakan senjata api (senpi) - Pencurian dengan kekerasan menggunakan senjata tajam (sajam)
6.	Kejahatan terhadap hak milik/barang	- Pencurian - Pencurian dengan pemberatan - Pencurian kendaraan bermotor - Pengrusakan/penghancuran barang - Pembakaran dengan sengaja - Penadahan
7.	Kejahatan terkait narkoba	Narkoba dan psikotropika
8.	Kejahatan terkait penipuan, penggelapan dan korupsi	- Penipuan/perbuatan curang - Penggelapan - Korupsi
9.	Kejahatan terhadap ketertiban umum	- Terhadap ketertiban umum

TABEL I: Klasifikasi Kejahatan
Sumber: BPS, 2023

Mencermati celah tersebut, tim peneliti menilai bahwa upaya deteksi aksi kejahatan secara manual tersebut tidak efisien. Untuk mengatasi masalah ini, tim peneliti mengusulkan pengembangan sebuah kerangka sistem yang dapat mendeteksi aksi kejahatan secara otomatis dari video CCTV. Dalam pengembangannya, tim peneliti membangun sebuah *dataset* berupa sekumpulan rekaman video CCTV aksi kejahatan di Indonesia yang diperoleh dari YouTube. Kemudian, *dataset* tersebut digunakan untuk mengembangkan sebuah model berbasis *anomaly detection* dengan dua pendekatan yang berbeda. Dengan kedua pendekatan ini, sistem deteksi diharapkan mampu mendeteksi aksi kejahatan dari video kamera pengawas berdasarkan nilai *anomaly score* pada setiap *frame*. Sistem ini dapat dipadukan dengan sebuah *alarm notifier* yang dapat memberi peringatan kepada petugas pengawas keamanan tentang kemungkinan aksi kriminal ketika nilai *anomaly score* cukup tinggi. Dengan demikian, petugas pengawas keamanan dapat segera menindaklanjuti dan memverifikasi keberadaan aksi kejahatan tersebut. Sistem deteksi ini diharapkan dapat mendeteksi aksi kejahatan secara *real time* agar kedepannya dapat diimplementasikan pada CCTV. Pengembangan sistem ini juga diharapkan dapat mengurangi *human error* dalam deteksi kejahatan dan membantu petugas pengawas keamanan dalam melakukan pengawasan yang lebih efisien.

B. Pertanyaan Penelitian

Berdasarkan latar belakang yang telah dijelaskan, pertanyaan penelitian yang akan tim peneliti ajukan adalah sebagai berikut.

1) Bagaimana efektivitas pendekatan *Multiple Instance Learning* dan *Video Vision Transformers* dengan *Sliding Sindow* dalam mendeteksi tindakan kriminal dalam video kamera pengawas?
2) Apa tantangan utama dalam implementasi pendekatan *Multiple Instance Learning* dan *Video Vision Transformers* dengan *Sliding Sindow* untuk mendeteksi tindakan kriminal dalam video kamera pengawas?
3) Bagaimana performa pendekatan *Multiple Instance Learning* dan *Video Vision Transformers* dengan *Sliding*

Sindow jika dibandingkan dengan pendekatan ekstraksi fitur menggunakan model berbasis *Convolutional Two-Stream Network*?
4) Bagaimana sistem deteksi tindakan kriminal dalam video pengawasan ini dapat diintegrasikan dengan sistem keamanan yang ada untuk membantu meningkatkan efisiensi dalam mendeteksi tindakan kriminal?

C. Tujuan Penelitian

Berikut adalah tujuan yang ingin dicapai dari penelitian ini.
1) Mengetahui bagaimana efektivitas pendekatan *Multiple Instance Learning* dan *Video Vision Transformers* dengan *Sliding Sindow* dalam mendeteksi tindakan kriminal dalam video kamera pengawas
2) Mengidentifikasi tantangan utama dalam implementasi pendekatan *Multiple Instance Learning* dan *Video Vision Transformers* dengan *Sliding Sindow* untuk mendeteksi tindakan kriminal dalam video kamera pengawas.
3) Mengetahui bagaimana performa pendekatan *Multiple Instance Learning* dan *Video Vision Transformers* dengan *Sliding Sindow* jika dibandingkan dengan pendekatan ekstraksi fitur menggunakan model berbasis *Convolutional Two-Stream Network*.
4) Mengetahui bagaimana sistem deteksi tindakan kriminal ini dapat diintegrasikan dengan sistem keamanan yang ada untuk membantu meningkatkan efisiensi dalam mendeteksi tindakan kriminal.

D. Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat bagi pihak-pihak yang membutuhkan, baik secara teoretis maupun praktis, diantaranya sebagai berikut.
1) Manfaat Teoretis
   - Sebagai *blueprint* untuk membuat sistem keamanan yang efisien di fasilitas publik dan dapat menjadi landasan bagi peneliti lain dalam melakukan penelitian yang sejenis
   - Penelitian ini diharapkan dapat menambah wawasan dan pengetahuan pembaca terkait deteksi kejadian pada video.
2) Manfaat Praktis
   - Membantu pemerintah dalam upaya untuk meningkatkan pelayanan fasilitas publik dengan meningkatkan keamanan pada fasilitas publik
   - Membantu meningkatkan rasa kenyamanan dan keamanan masyarakat pada fasilitas publik.

E. Batasan Penelitian

Batasan penelitian pada metode yang dipaparkan dalam makalah adalah sebagai berikut
1) Jenis tindakan kriminal yang dapat dideteksi hanya penyiksaan (*abuse*), penangkapan (*arrest*), pembakaran (*arson*), penyerangan (*assault*), kecelakaan di jalan raya (*road accident*), pencurian (*burglary*), ledakan (*explosion*), perkelahian (*fighting*), perampokan (*stealing*), penembakan (*shooting*), pencurian di toko (*shoplifting*), dan vandalisme (*vandalism*)
2) Penelitian hanya berfokus pada pembuatan model deteksi.

II. Kajian Pustaka

A. Weakly Supervised dan Multiple Instance Learning

Deteksi anomali adalah proses untuk menemukan data atau objek yang tidak biasa dalam kumpulan data. Tujuannya adalah untuk mengidentifikasi kejadian atau observasi yang berbeda dari pola mayoritas data [5]. Deteksi anomali digunakan dalam berbagai bidang, seperti mendeteksi penipuan, penyakit, dan tindak kejahatan. Seringkali, deteksi anomali dilakukan tanpa label data yang lengkap dan akurat. Hal ini juga kerap terjadi dalam deteksi anomali pada video, di mana pemberian label pada setiap *frame* video dapat sangat mahal.

Weakly Supervised Learning adalah cara melatih model *machine learning* di mana informasi label data tidak sempurna [6]. Terdapat tiga jenis *Weakly Supervised Learning* : *Incomplete Supervision* (label tidak lengkap), *Inexact Supervision* (label tidak rinci), dan *Inaccurate Supervision* (label tidak selalu benar) [6]. Dalam pengembangan model deteksi anomali video, pendekatan *Inexact Supervision* dapat mengurangi biaya pelabelan yang mahal.

Salah satu cara menerapkan *Inexact Supervision* adalah dengan teknik *Multiple Instance Learning*. Teknik ini menggunakan label dari kelompok data yang lebih besar (*bag*) untuk memberi label pada data yang lebih kecil (*instance*) [7]. Dalam deteksi anomali, *bag* dianggap memiliki anomali jika terdapat *instance* atau *frame* di dalamnya yang mengandung anomali. Pendekatan ini terbukti efektif dalam beberapa penerapan model deteksi anomali [8].

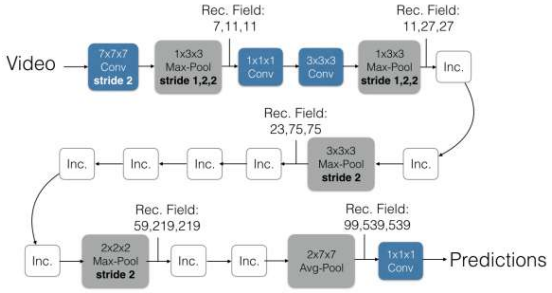
B. I3D Two Stream

I3D (*Inflated 3D ConvNet*) *Two Stream* adalah model jaringan saraf tiruan yang ditujukan untuk menganalisis video. Model ini merupakan hasil pengembangan dari model 2D ConvNet yang diperluas (*inflated*) menjadi tiga dimensi untuk mengolah data video yang memiliki informasi temporal dan spasial secara lebih efektif. Dibandingkan dengan model 2D ConvNet, model ini memiliki dimensi yang lebih kompleks dan memiliki lebih banyak parameter, yang artinya performa model akan jauh lebih baik dibandingkan model 2D ConvNet karena terdapat pemrosesan RGB dan pemrosesan data aliran optik yang lebih canggih [9].

Dalam menganalisis suatu video, *I3D Two Stream* model memiliki beberapa langkah. Diagram (Gambar 1) menjelaskan proses yang dilakukan oleh *I3D Two Stream model*.

Pertama, model menerima dua jenis input, yaitu *RGB Stream* (model mengambil informasi terkait warna dan detail spasial dari setiap *frame*) dan *Optical Flow Stream* (model mengambil informasi terkait hasil analisis perbedaan antara *frame* berturut-turut untuk menangkap gerakan). Kedua jenis input ini kemudian diproses melalui proses konvolusi 3D yang diinflasi dari model konvolusi 2D. Terdapat dua langkah pada bagian ini, yaitu pertama filter konvolusi 2D diinflasi

ISSN: xxxxxx



Gambar 1: Cara Kerja Model *I3D Two Stream*

menjadi 3D dengan cara melakukan filter secara iterasi untuk menangkap pola gerakan. Selanjutnya, *pooling layer* juga diinflasi menjadi 3D untuk melakukan *downsampling* pada dimensi temporal dan spasial. Pooling 3D dilakukan untuk mengurangi ukuran data dan tetap mempertahankan informasi penting pada video tersebut.

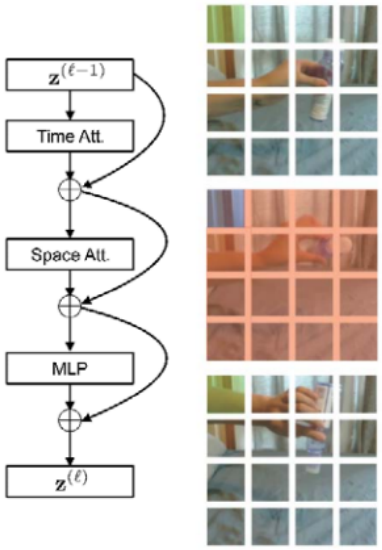
Selanjutnya, setiap *stream* tersebut akan dilakukan *feature extraction*. Pada *RGB stream*, model mengekstrak fitur spasial dari setiap *frame*. Sementara itu, pada *Optical Flow stream*, model mengekstrak fitur gerakan dari perubahan antar *frame*.

Setelah dilakukan *feature extraction* dari kedua stream, hasilnya akan digabungkan untuk dilakukan analisis lebih lanjut. Langkah berikutnya adalah model menggunakan lapisan *fully connected* untuk melakukan klasifikasi atau deteksi berdasarkan fitur yang digabungkan. Misalnya, deteksi pengenalan aksi/gerakan, deteksi anomali, dan lain-lain.

C. Video Vision Transformers

Vision Transformers (VTs) merupakan salah satu cabang dari arsitektur model *Transformers*. Model ini pertama kali diperkenalkan oleh Dosovitskiy et al. dalam makalah "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" pada tahun 2020 [10]. Model VTs berasal dari pengembangan arsitektur model *Transformers*, yang awalnya kerap digunakan pada bidang pemrosesan bahasa alami (NLP), untuk tugas-tugas pengenalan citra. Dengan berbasis *Transformers*, VTs memiliki perbedaan pendekatan dalam menangkap hubungan spasial dalam gambar melalui mekanisme *self-attention*, yang tidak dimiliki oleh pendekatan konvolusi yang digunakan dalam Convolutional Neural Networks (CNNs) [11].

Dalam perluasan penggunaan VTs ke dalam tugas pengenalan citra video, beberapa penyesuaian telah dilakukan untuk memproses dimensi temporal dalam konten video. Model seperti VideoBERT [12] dan TimeSformer [13] telah menunjukkan hasil yang menjanjikan dalam memahami konten video dengan menggabungkan informasi spasial dan temporal. VideoBERT mengadopsi pendekatan BERT (*Bidirectional Encoder Representations from Transformers*) untuk memproses urutan *frame* video sebagai input teks untuk menunjukkan kemampuannya dalam memahami aksi dan konteks dalam video tanpa memerlukan anotasi. Sementara itu, TimeSformer memanfaatkan *self-attention* secara spasial dan temporal untuk mengolah urutan *frame* video, sehingga menghasilkan performa yang lebih baik dibandingkan model CNN.



Gambar 2: Cara Kerja Timesformer

III. METODOLOGI

A. Deskripsi Solusi

Untuk mengatasi permasalahan tindakan kriminal di lingkungan masyarakat, tim peneliti merancang sebuah sistem deteksi tindakan kriminal yang dapat diintegrasikan dengan kamera pengawas (CCTV). Sistem ini bertujuan untuk mengurangi potensi kelalaian atau *human error* yang kerap terjadi dalam pengawasan CCTV konvensional. Dengan memanfaatkan pendekatan *Multiple Instance Learning* dan *Video Vision Transformers*, sistem ini mampu menganalisis setiap *frame* video dari kamera pengawas. Ketika tingkat ketidaknormalan atau *anomaly score* yang terdeteksi cukup tinggi atau melebihi ambang batas tertentu, sistem dapat memberikan alarm sebagai peringatan dini. Ambang batas tersebut dapat disesuaikan untuk setiap tempat yang akan diawasi. Sebagai contoh, tempat yang membutuhkan pengamanan ketat dapat menggunakan *threshold* yang lebih rendah untuk meminimalisir *false negative*. Diharapkan, melalui implementasi sistem deteksi ini, tindakan kriminal dapat lebih cepat teridentifikasi dan ditangani oleh pihak berwenang, sehingga meningkatkan keamanan dan keselamatan masyarakat.

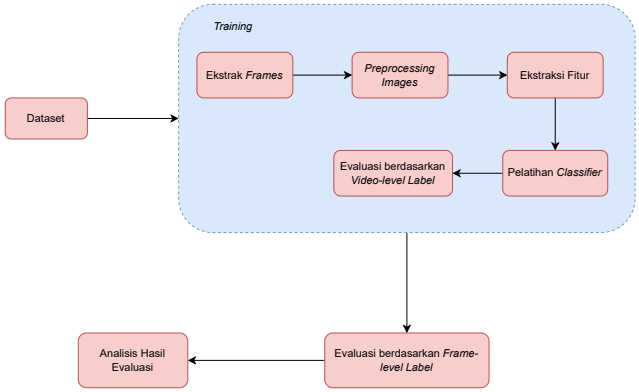
B. Dataset

Penelitian ini menggunakan *dataset* UCF Crime [14], yang berisi 1900 video dengan durasi total 128 jam. *Dataset* ini mencakup 13 jenis tindakan kriminal: penyiksaan (*abuse*), penangkapan (*arrest*), pembakaran (*arson*), penyerangan (*assault*), kecelakaan di jalan raya (*road accident*), pencurian (*burglary*), ledakan (*explosion*), perkelahian (*fighting*), perampokan (*stealing*), penembakan (*shooting*), pencurian di toko (*shoplifting*), vandalisme (*vandalism*), serta video tanpa

tindakan kriminal. Semua video dengan tindakan kriminal dikelompokkan menjadi satu grup anomali, sedangkan video tanpa tindakan kriminal dikelompokkan secara terpisah. Selain *dataset* tersebut, penelitian ini juga memanfaatkan *dataset* tambahan yang dikumpulkan sendiri oleh tim peneliti. *Dataset* tambahan ini berisi rekaman video kamera pengawas di Indonesia, dengan masing-masing jenis tindakan kriminal diwakili oleh 2 video. Dalam penelitian ini, *dataset* UCF Crime digunakan untuk melatih (*training*) dan menguji (*testing*) model dengan label tingkat klip video. Sementara itu, *dataset* tambahan digunakan untuk mengevaluasi performa model dalam mendeteksi anomali pada setiap *frame* sehingga pemberian label pada setiap *frame* dilakukan secara manual.

C. Metode Penelitian

Secara umum, berikut adalah alur dari setiap metode yang digunakan pada penelitian ini (Gambar 3).



Gambar 3: Alur Penelitian

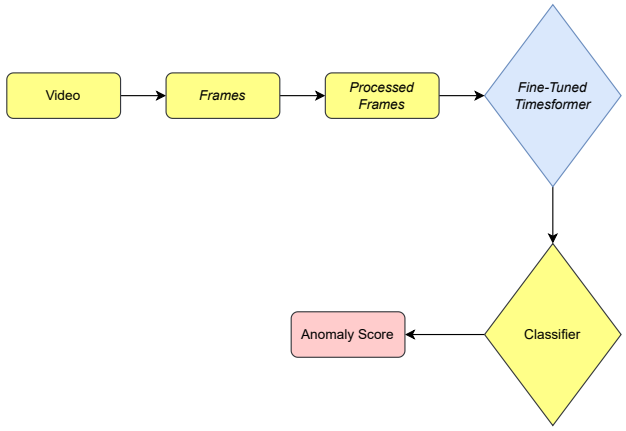
1) **Metode *Inflated 3D Convolution Networks Two Stream***
Langkah pertama yang dilakukan adalah mengekstraksi 30 *frame* per detik dari setiap video dalam *dataset*. Setiap *frame* kemudian diatur sehingga ukurannya menjadi 320x240 piksel dan dihasilkan *frame optical flow* yang dapat menangkap gerakan antar *frame*. Selanjutnya, fitur dari *RGB stream* dan *optical flow stream* diekstraksi menggunakan model ResNet101 yang telah dilatih sebelumnya pada *dataset* pengenalan aksi Kinetics-400 [15]. Setelah itu, hasil ekstraksi tersebut diterapkan normalisasi.
Hasil ekstraksi fitur kemudian digunakan untuk melatih model *classifier*. Model yang digunakan berupa jaringan saraf tiruan (*neural network*) dengan 3 lapisan linear, fungsi aktivasi ReLU, dan *dropout* pada setiap *layer* kecuali *layer* terakhir yang menggunakan fungsi *sigmoid* untuk menghasilkan nilai probabilitas. Model dilatih menggunakan teknik *Multiple Instance Learning* dimana label tingkat video (*video-level label*) digunakan saat pelatihan.
Model kemudian dievaluasi menggunakan label tingkat video dari bagian uji *dataset* UCF Crime dan label

Gambar 4: Cara Kerja Metode *I3D Two Stream*

tingkat *frame* dari *dataset* tambahan yang tim peneliti kumpulkan. Dalam mengukur performa model, tim peneliti menggunakan metrik ROC AUC [16] untuk mengukur performa model. Selanjutnya, dilakukan analisis performa model dari segi ketepatan dan waktu inferensi.

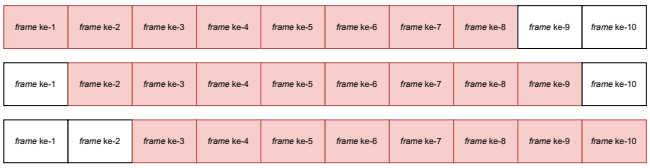
2) **Metode *Video Vision Transformers dengan Sliding Window***
Berbeda dengan pendekatan sebelumnya yang menggunakan model berbasis *Convolutional Networks*, pada metode ini tim peneliti mengusulkan pendekatan berbeda berbasis model *Transformers*. Pada penelitian ini, digunakan model *TimeSformer* [13] yang merupakan salah satu model *Video Vision Transformers* [17] yang telah dilatih pada *dataset* pengenalan aksi, seperti Kinetics-400 [15]. *TimeSformer* pada dasarnya hanya memiliki kemampuan untuk mengklasifikasikan satu klip video ke dalam kategori aksi tertentu.
Dalam metode ini, proses pelatihan model melibatkan ekstraksi sampel *frame* dari setiap video. Secara acak, 8 *frame* dipilih dari setiap video. Kemudian setiap *frame* diterapkan *preprocessing*, di antaranya mengubah ukuran menjadi 224x224 piksel, menskalakan nilai piksel ke rentang 0 hingga 1, dan melakukan normalisasi nilai piksel.
Frame yang telah diproses digunakan untuk melakukan *fine-tuning* pada model *TimeSformer* dan melatih model *classifier* sederhana berupa satu *neuron linear* dengan fungsi aktivasi *sigmoid*. Pelatihan model dilakukan menggunakan teknik *Multiple Instance Learning* dengan label tingkat video (*video-level label*).
Model ini kemudian diuji menggunakan label tingkat video (*video-level label*) dari bagian uji dataset UCF Crime dan label tingkat *frame* (*frame-level label*) dari *dataset* tambahan. Untuk memprediksi anomali pada setiap *frame*, digunakan pendekatan *Sliding Window*, di mana 8 *frame* dikumpulkan dan digunakan untuk mem-

ISSN: xxxxxx



Gambar 5: Cara Kerja Metode *Vision Transformers*

prediksi *anomaly score*. *Frame* selanjutnya diprediksi dengan menggunakan *frame* pertama hingga kesembilan, dan seterusnya. Analisis performa juga dilakukan pada metode ini.



Gambar 6: Visualisasi *Sliding Window*

D. Metrik Evaluasi

Penelitian ini menggunakan metrik *Area Under the Receiver Operating Characteristic Curve* (AUC-ROC) [16] untuk mengukur performa model klasifikasi pada berbagai ambang batas (*threshold*). *Receiver Operating Characteristic* (ROC) adalah kurva probabilitas yang menggambarkan hubungan antara *true positive rate* (TPR) dan *false positive rate* (FPR). AUC adalah luas area di bawah kurva ROC, yang mengindikasikan seberapa baik model dapat membedakan antara kelas positif dan negatif. Semakin tinggi nilai AUC, semakin baik kemampuan model dalam membedakan kedua kelas tersebut. Metrik ini kemudian digunakan untuk mengukur ketepatan prediksi setiap metode dalam mendeteksi anomali pada setiap frame (*frame-level*) dan ada atau tidaknya anomali pada suatu video (*video-level*).

IV. Hasil Eksperimen dan Pengujian

Dalam penelitian ini, metode berbasis jaringan konvolusi dilatih menggunakan *optimizer Adam* dengan *learning rate* 0.001 selama 75 *epochs* dan *batch size* 30, serta menggunakan *scheduler* MultiStepLR. Sementara itu, metode berbasis *Transformers* dilatih selama 3 *epochs* dengan *optimizer Adam* yang memiliki *learning rate* 0.00005 dan *batch size* 2, serta menggunakan *scheduler linear*. Selain itu, pada penelitian ini digunakan *loss function* berupa *binary cross entropy*.

Berikut adalah ini adalah hasil pengujian yang tim peneliti dapatkan ketika kedua metode diujikan pada *video-level label* dam *frame-level label*.

Metode	AUC-ROC Score
Metode *I3D Two Stream*	0.841
Metode *Video Vision Transformers*	0.942

TABEL II: AUC ROC *score video-level label*

Metode	Rata-rata AUC-ROC Score
Metode *I3D Two Stream*	0.515
Metode *Video Vision Transformers*	0.749

TABEL III: AUC ROC *score frame-level label*

Terlihat bahwa metode *Video Vision Transformers* memiliki nilai metrik yang lebih baik dibandingkan metode *I3D Two Stream* baik pada pengujian menggunakan *video-level label* ataupun *frame-level label*.

Selain itu, tim peneliti juga melakukan pengujian terhadap kecepatan inferensi dari setiap metode. Untuk itu, dilakukan perhitungan rata-rata banyak *frame* yang dapat diproses per detiknya oleh masing-masing metode.

Metode	***Frame*** tiap detik
Metode *I3D Two Stream*	11.1
Metode *Video Vision Transformers*	5.6

TABEL IV: AUC ROC Score *frame-level label*

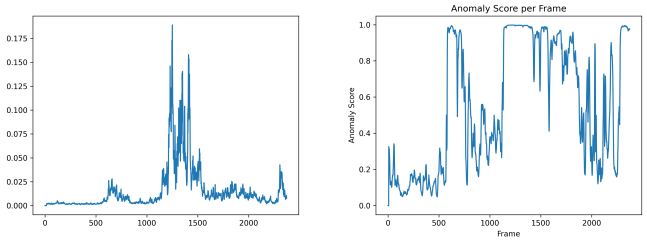
V. Analisis Hasil Eksperimen dan Pengujian

Hasil pengujian pada Tabel II menunjukkan bahwa metode *Video Vision Transformers* (AUC-ROC score 0.942) mengungguli metode *I3D Two Stream* (AUC-ROC score 0.841) dalam hal ketepatan prediksi pada tingkat video. Hal ini mengindikasikan bahwa model *Video Vision Transformers* lebih mampu membedakan antara video yang mengandung tindakan kriminal dan video normal secara keseluruhan.

Selanjutnya, pada pengujian dengan label tingkat frame (Tabel IV), metode *Video Vision Transformers* kembali menunjukkan performa yang lebih baik dengan rata-rata AUC-ROC score 0.749, dibandingkan dengan metode *I3D Two Stream* yang hanya mencapai 0.515. Keunggulan ini menunjukkan bahwa *Video Vision Transformers* lebih efektif dalam mengidentifikasi tindakan kriminal pada tingkat *frame* individual, yang berarti model ini lebih akurat dalam melokalisasi momen-momen spesifik terjadinya tindakan kriminal dalam sebuah video.

Secara keseluruhan, hasil evaluasi ini menunjukkan potensi metode *Video Vision Transformers* sebagai pendekatan yang lebih unggul dalam pengembangan sistem deteksi tindakan kriminal berbasis video, terutama dalam hal ketepatan. Namun, perlu diperhatikan bahwa metode ini memiliki kelemahan dalam efisiensi, terlihat dari jumlah *frame* yang dapat diproses per detik yang lebih sedikit dibandingkan metode *I3D Two Stream*.

Selain perbedaan performa dan efisiensi, metode *I3D Two Stream* juga memiliki kelemahan dalam menghasilkan model dengan tingkat kepercayaan prediksi yang tinggi. Hal ini terlihat pada Gambar 8, yang memvisualisasikan *anomaly score* dari salah satu video dalam *dataset*. Video tersebut memiliki



Gambar 7: Anomaly Score Metode *I3D* (Kiri) dan *Transformers* (Kanan)

dua interval *frame* yang mengandung anomali, yaitu antara *frame* 623 hingga 794 dan *frame* 1185 hingga 1485. Meskipun kedua model cenderung dapat mendeteksi peristiwa anomali kedua, model *I3D Two Stream* gagal mengidentifikasi peristiwa anomali pertama. Selain itu, terlihat bahwa *anomaly score* yang dihasilkan oleh model *I3D Two Stream* secara umum lebih rendah dibandingkan model *Video Vision Transformers*, bahkan ketika terjadi peristiwa anomali. Temuan seperti ini juga ditemukan pada video-video lain pada *dataset*. Perbedaan signifikan dalam *anomaly score* ini menunjukkan bahwa model *I3D Two Stream* kurang mampu membedakan secara tegas antara *frame* normal dan *frame* yang mengandung anomali.

Gambar 8: Gambaran Prediksi Metode *I3D* (Kiri) dan *Transformers* (Kanan)

VI. KESIMPULAN

Penelitian ini menunjukkan bahwa pendekatan *Multiple Instance Learning* dan *Video Vision Transformers* dengan *Sliding Window* cukup efektif dalam mendeteksi tindakan kriminal. Pendekatan ini juga terbukti lebih unggul dalam hal ketepatan dibandingkan pendekatan berbasis jaringan konvolusi, walaupun masih terdapat tantangan dalam hal efisiensi pemrosesan *frame*. Oleh karena itu, penting untuk diadakannya sumber daya komputasi yang mumpuni ketika teknologi ini diintegrasikan di lapangan. Selain itu, dalam penerapannya, penting untuk menyesuaikan *threshold anomaly score* berdasarkan urgensi setiap lokasi yang diawasi untuk menghasilkan *alarm notifier* yang optimal. Sebagai penutup, diharapkan penelitian dapat bermanfaat untuk pengembangan dan pengelolaan sistem keamanan di masa mendatang.

REFERENSI

[1] A. G. Perry and P. A. Potter, "Buku ajar fundamental keperawatan vol, 2." Egc, 2005.
[2] Kepolisian Negara Republik Indonesia (POLRI). (2024) Data kejahatan. [Online]. Available: https://pusiknas.polri.go.id/data_kejahatan
[3] Badan Pusat Statistik, "Statistik kriminal 2023. vol. 14," https://www.bps.go.id/id/publication/2023/12/12/5edba2b0fe5429a0f232c736/statistik-kriminal-2023.html, 2023, online; accessed 13 May 2024.
[4] G. R. C. K. BP, "Peran kamera pengawas closed-circuit television (cctv) dalam kontra terorisme," *Jurnal Lemhannas RI*, vol. 9, no. 4, pp. 100–116, 2021.
[5] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009.
[6] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National science review*, vol. 5, no. 1, pp. 44–53, 2018.
[7] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognition*, vol. 77, pp. 329–353, 2018.
[8] Y. Zhou, Y. Qu, X. Xu, F. Shen, J. Song, and H. Shen, "Batchnorm-based weakly supervised video anomaly detection," *arXiv preprint arXiv:2311.15367*, 2023.
[9] Z. Wang, "Anomaly detection in surveillance videos based on two-stream inflated 3d conv net and weakly supervised learning," in *CIBDA 2022; 3rd International Conference on Computer Information and Big Data Applications*, 2022, pp. 1–5.
[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
[11] J. Bi, Z. Zhu, and Q. Meng, "Transformer in computer vision," in *2021 IEEE International conference on computer science, electronic information engineering and intelligent control technology (CEI)*. IEEE, 2021, pp. 178–188.
[12] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "Videobert: A joint model for video and language representation learning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7464–7473.
[13] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *ICML*, vol. 2, no. 3, 2021, p. 4.
[14] Waqas Sultani, Chen Chen, Mubarak Shah. (2018) Real-world anomaly detection in surveillance videos. [Online]. Available: https://www.crcv.ucf.edu/projects/real-world/
[15] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.
[16] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320396001422
[17] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6836–6846.