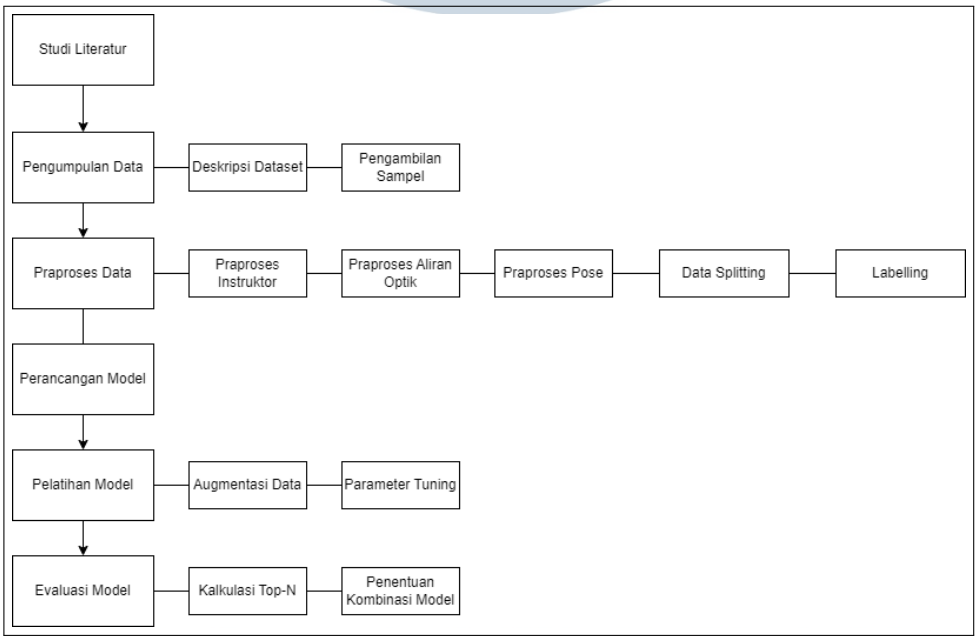
BAB 3
METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Penelitian ini berfokus dalam memprediksi bahasa isyarat Indonesia (BISINDO) berupa video menjadi kata. Prediksi bahasa isyarat dilakukan dengan menggunakan metode yang dilakukan pada penelitian sebelumnya dengan menggunakan 3 jenis model input (rgb, aliran optik, dan pose). Hasil prediksi akan diuji menggunakan metode top-N untuk mencari tingkat akurasi. Analisis akurasi akan dilakukan pada akhir penelitian untuk menjawab rumusan masalah serta tujuan penelitian yang telah dilakukan.

3.2 Metode Penelitian

Berikut merupakan diagram alur jalannya penelitian. Pada bab ini akan dijelaskan alur-alur yang telah dijalankan:



Gambar 3.1. Alur jalannya penelitian

3.3 Studi Literatur

Karena keterbatasan penelitian BISINDO pada level kata, referensi terhadap bahasa isyarat lain (seperti ASL dan KSL) dapat memberikan inspirasi pendekatan metode maupun pembentukan dataset yang lebih besar. Sehingga Studi literatur dilakukan dengan mengidentifikasi 3 penelitian pada bahasa isyarat diluar BISINDO dan 2 penelitian bahasa isyarat BISINDO dikarenakan bahasa isyarat mempunyai sifat dasar yang sama. Berikut merupakan penelitian-penelitian yang telah dilakukan terkait deteksi bahasa isyarat pada level kata:

Tabel 3.1. Studi terdahulu terhadap mesin prediksi bahasa isyarat berlevel kata

Penelitian	**Model / Metode**	**Objek**	bahasa
Andi et al [9]	CNN & LSTM	BISINDO	Indonesia
Handhika et al [10]	Hidden-Markov Model (HMM)	BISINDO	Indonesia
maruyama et al [1]	I3D & ST-GCN	WLASL[5] dan MS-ASL[6]	Inggris (Amerika)
Ogulcan Ozdemir et al [3]	DeepHand, STGCN, DAN, dan MC-LSTM	AUTSL[8] dan BosphorusSign22k [7]	Turki
Shin et al [4]	GCN & CNN	KSL	Korea

3.3.1 Penelitian Terkait BISINDO pada level kata

Umumnya, penelitian BISINDO hanya terdiri pada level alfabet, sehingga Penelitian mengenai prediksi BISINDO pada level kata masih sangat terbatas.

Penelitian andi et al [9] memprediksi 8 kata + 2 huruf BISINDO menggunakan metode CNN dan LSTM dengan menggunakan video dengan latar belakang yang telah dihapus dan melalui proses *gaussian blur*. Hasil penelitian memperoleh akurasi tertinggi sebesar 96% dengan kombinasi model CNN yang diteruskan ke LSTM

Penelitian Handhika et al [10] memprediksi 25 kata menggunakan metode Hidden-Markov Model (HMM) dengan menggunakan data pose tubuh yang didapat dari video. Hasil penelitian memperoleh akurasi sebesar 65%.

Kedua penelitian tersebut hanya memakai satu *channel* video global tanpa menggunakan pemotongan instruktor bahasa isyarat sehingga memungkinkan

terjadinya kendala berupa *noise* yang terdiri dari latar belakang yang tidak relevan dalam prediksi bahasa isyarat.

3.3.2 Penelitian Terkait Bahasa Isyarat Bahasa Asing

Sedangkan untuk penelitian terkait prediksi bahasa isyarat selain bahasa indonesia adalah sebagai berikut.

Maruyama et al [1] memprediksi bahasa isyarat amerika yang terdiri dari WLASL dan MS-ASL. Berbeda dengan penelitian BISINDO yang hanya memakai satu fitur, penelitian ini menggunakan berbagai fitur spatial yang terdapat pada sebuah video, fitur ini terbagi menjadi 3 bagian besar, yaitu fitur *base stream* yang terdiri dari aliran optik dan video rgb instruktor, fitur *local image stream* yang terdiri dari video potongan tangan dan muka, dan *skeleton stream* yang terdiri dari pose tubuh. Masing-masing fitur dideteksi oleh model I3D, untuk *skeleton stream* dideteksi oleh model ST-GCN. Seluruh fitur akan dirata-ratakan sehingga memperoleh prediksi akhir. Hasil penelitian mempunyai akurasi yang tinggi sebesar 81.38% pada subset WLASL100 yang terdiri dari 100 kata. Akurasi penelitian ini lebih tinggi dari penelitian sebelumnya yang hanya memprediksi video rgb global menggunakan model I3D.

Ogulcan ozdemir et al [3] memprediksi bahasa isyarat turki yang terdiri dari 2 dataset AUTSL dan BosphorusSign22k. Mirip seperti penelitian Maruyama et al yang menggabungkan fitur lokal dengan fitur global, penelitian ini menggunakan gabungan antara CNN dan RNN dimana terdapat 3 fitur dengan model prediksi yang berbeda: tangan (DeepHand), muka (DAN), dan pose (ST-GCN) yang diprediksi dan masing-masing prediksi akan memasuki jaringan RNN berupa MC-LSTM. Masing-masing hasil prediksi akan dirata-ratakan sehingga memperolah prediksi akhir. Hasil akurasi tertinggi pada penelitian ini adalah sebesar 92.58% pada dataset BosphorusSign22k.

Shin et al [4] memprediksi bahasa isyarat korea yang terdiri dari dataset KSL. Sama seperti penelitian yang dilakukan Handhika et al, penelitian ini menggunakan data pose tubuh, namun dengan menggunakan model jaringan graf konvolutional dan CNN dimana penelitian ini akan memprediksi titik-titik pose tubuh dan hubungan antara titik-titik pose tubuh. Hasil kedua prediksi akan digabungkan dan diberikan ke jaringan CNN sehingga memperoleh prediksi akhir. Akurasi pada model ini mencapai 99.87%

3.3.3 Analisis Perbandingan dan Celah

Jika dibandingkan dengan penelitian-penelitian di atas, pendekatan terhadap BISINDO masih sangat terbatas baik dari segi jumlah kata, keragaman input, maupun kompleksitas model. Jika dilihat dari penelitian-penelitian yang dilakukan pada bahasa isyarat lain, penelitian-penelitian tersebut menggabungkan beberapa fitur yang diprediksi oleh model-model yang terpisah yang nantinya akan digabungkan untuk memperoleh prediksi akhir. Hal ini dapat dijelaskan melalui penelitian yang dilakukan Maruyama et al dimana akurasi prediksi model terhadap beberapa fitur pada sebuah video bahasa isyarat akan lebih tinggi daripada model yang hanya memprediksi satu fitur saja. Sehingga prediksi BISINDO masih bisa ditingkatkan lebih lanjut akurasinya menggunakan metode ini.

3.4 Pengumpulan Data

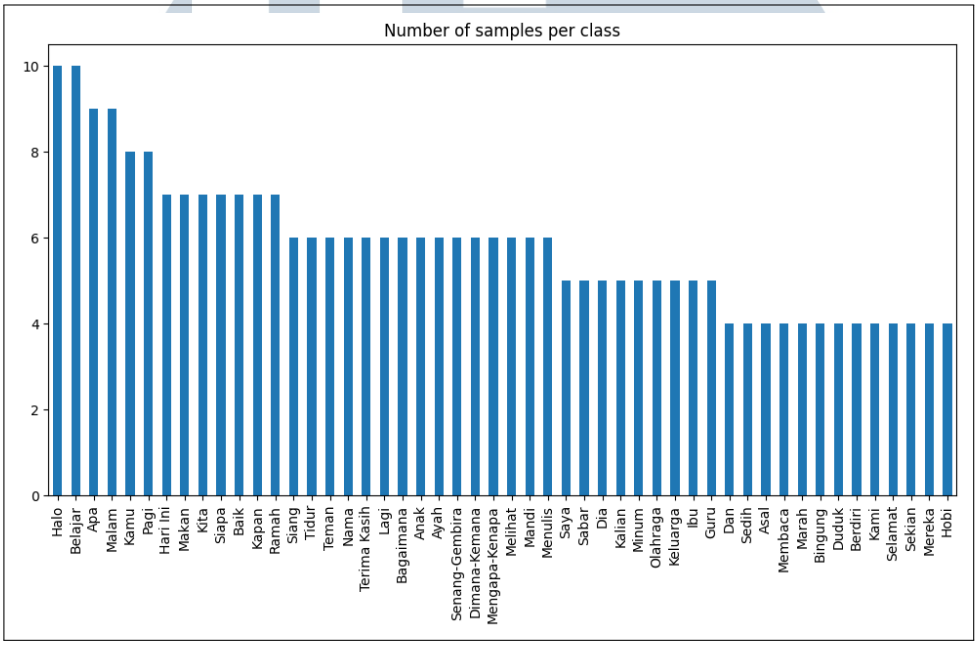
3.4.1 Deskripsi Dataset

Dataset yang digunakan adalah dataset BISINDO [47] yang dapat didownload di kaggle. Dataset terdiri dari video .mp4 dan .mov berupa huruf (26 alfabet) dan kata (50 leksikon) bahasa isyarat Indonesia yang bersumber dari Pusat Layanan Juru Bahasa Isyarat (PLJ) dan kontribusi penulis dan dataset sekunder berupa video akses bebas yang terdapat dari Youtube dan Kaggle. Isi dari dataset awal terdiri dari video original yang telah di praproses sehingga terdapat video yang di *mirror* (pembalikan secara horizontal), resolusi rendah dan tinggi, video durasi rendah dan tinggi, serta *framerate* yang beragam. Terdapat beberapa video yang mempunyai ekstensi .mov dan .mp4, ekstensi .mov seringkali mempunyai kualitas yang lebih baik daripada ekstensi .mp4 [48], namun mempunyai ukuran file yang lebih besar.

3.4.2 Pengambilan Sampel

Dikarenakan topik pada penelitian yang membataskan prediksi bahasa isyarat hanya pada level kata, penelitian ini tidak menggunakan dataset huruf dan akan mengambil seluruh kata (50 leksikon). Pemangkasan video pada dataset diperlukan untuk menghilangkan duplikasi pada dataset sehingga setiap video pada dataset hanya terdiri dari video yang unik dan berkualitas tinggi untuk diteruskan ke tahap praproses. Diasumsikan bahwa video yang mempunyai kualitas tinggi

merupakan video yang mempunyai ekstensi file .mov dan mempunyai durasi dan resolusi yang lebih besar dari versi video yang sama. Teknik pemangkasan ini tidak memperdulikan jika video telah di mirror dikarenakan makna dari bahasa isyarat tidak berubah jika video di-mirror. Hasil pemangkasan video terdiri dari 290 video yang mempunyai warna latar yang berbeda pada setiap sampel di sebuah kata.



Gambar 3.2. Jumlah sampel pada masing-masing kata

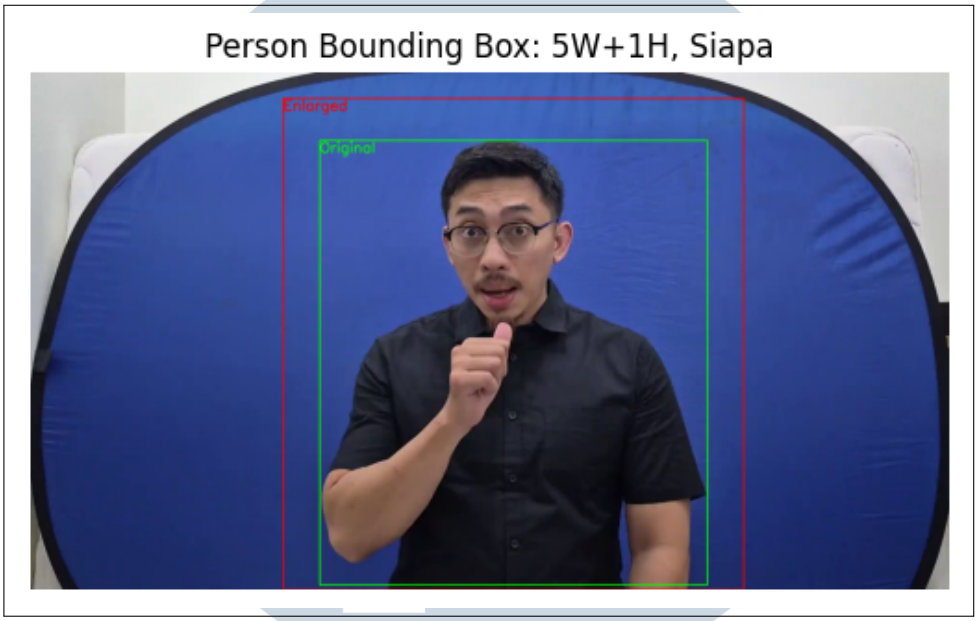
3.5 Praproses Data

Rangkaian teknik praproses data mengikuti penelitian yang maruyama et al terhadap praproses model I3D dan ST-GCN dimana terdapat praproses instruktor, aliran optik, dan praproses pose. File hasil praproses akan disimpan menggunakan *library* pickle dan di kompress menggunakan gzip.

3.5.1 praproses instruktor

Untuk praproses gambar pada video utama, pertama-tama, video akan memprediksi area dimana instruktor bahasa isyarat berada menggunakan YOLOv11 pada setiap *frame* video. Area ini akan diperbesar sebanyak $\sqrt{2}$ dan akan dipotong dan di ubah ukurannya menjadi 256x256 sehingga menjadi video instruktor bahasa isyarat yang berukuran 256x256 rgb. Ketika area pembesaran lebih besar daripada ukuran video, akan di *clamp* berdasarkan dimensi video. Berikut merupakan

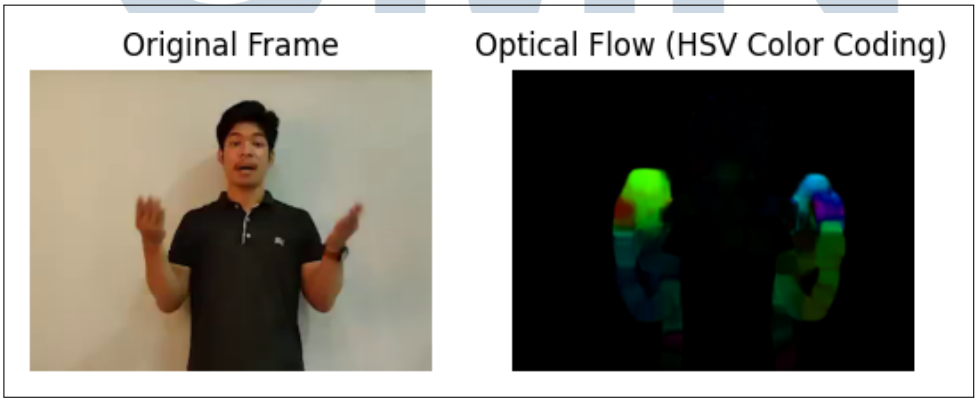
visualisasi gambar yang telah diprediksi menggunakan YOLOv11, kotak hijau merupakan prediksi YOLOv11 sedangkan kotak merah merupakan pembesaran ukuran sebesar $\sqrt{2}$.



Gambar 3.3. Metode praproses instruktor pada tahapan pertama

3.5.2 Praproses aliran optik

Praproses aliran optik terdiri dari praproses menggunakan algoritma TV-L1 dan Farneback menggunakan data hasil praproses instruktor sebelumnya. Dikarenakan algoritma aliran optik menghasilkan jumlah frame $M - 1$, maka frame terakhir akan di duplikasi sehingga menjaga ukuran frame.



Gambar 3.4. praproses algoritma aliran optik farneback dengan visualisasi menggunakan HSV

3.5.3 Praproses Pose

Praproses pose dilakukan dengan memprediksi seluruh frame pada video hasil praproses instruktor dengan menggunakan prediksi pose MediaPipe, hasil prediksi MediaPipe berupa koordinat xyz 33 titik pose tubuh dan 2x21 pose titik tangan kiri dan tangan kanan. Pada penelitian ini, 4 titik (indeks 14, 12, 11, dan 13) pose torso tubuh akan di ekstrak sedangkan 21x2 titik pose tangan akan dipakai seluruhnya. Titik-titik pose pada gambar akan diekstrak dan digabungkan dengan format $P_{pose} = (P_0^b ... P_4^b, P_5^{lh} ... P_{25}^{lh}, P_{26}^{rh} ... P_{46}^{rh})$ dimana P merupakan titik pose 3D, P^b merupakan titik pose tubuh, P^{lh} merupakan titik pose tangan kiri, dan P^{rh} merupakan titik pose tangan kanan. Ketika pose dari sebuah bagian tidak terdeteksi oleh MediaPipe, maka nilai seluruh titik pada pose tersebut akan diisi dengan 0.

3.5.4 Data Splitting

Data *splitting* merupakan pembagian dari dataset awal yang berjumlah 290 video menjadi 3 bagian *train*, validasi, dan *test*. Rasio pembagian dataset dilakukan sesuai dengan penelitian sebelumnya yang membagi dataset menjadi 4:1:1 masing-masing data *train*, validasi, dan *test*. Hasil data *split* menunjukan data *train* sebesar 193 video, validasi sebesar 48 video, dan *test* sebesar 49 video.

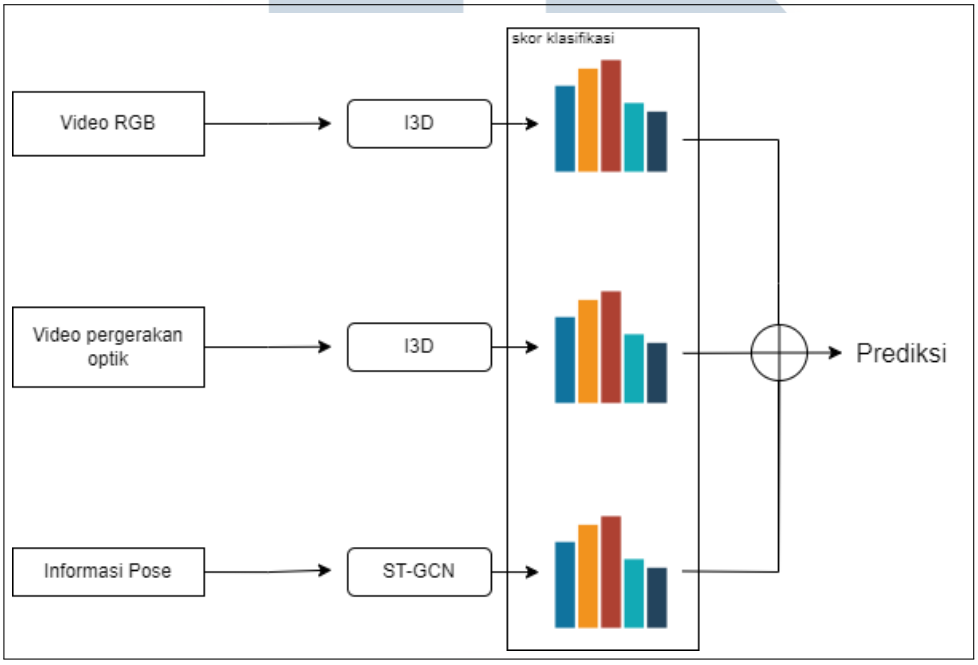
3.5.5 Labelling

Struktur dataset terdiri dari 2 *folder* dimana *folder* pertama terdiri dari frasa atau kelompok kata, frasa terdiri dari *folder* 5W+1H, Kata Ganti Orang, Kata Kerja, Kata Lainnya, dan Kata Sifat. di dalam *folder* frasa, terdapat kumpulan *folder* kata sesuai dengan gambar 3.2 yang didalamnya berisikan video. Sehingga untuk pengambilan label pada kumpulan video akan dilakukan berdasarkan masing-masing *folder* kata, dan tidak menggunakan *folder* frasa.

3.6 Perancangan Model

Terdapat 4 model yang dilatih secara terpisah berdasarkan data yang telah di praproses, model pertama merupakan model I3D yang menerima input data video rgb instruktor, model kedua dan ketiga merupakan model I3D yang menerima data aliran optik (TV-L1 dan farneback) dari video rgb instruktor, dan model terakhir merupakan model ST-GCN yang menerima input pose. Tiap-tiap aliran data pada

model akan melalui tahap augmentasi terlebih dahulu sebelum dilatih ke model untuk memastikan setiap data mempunyai dimensi temporal sebesar 64 frame, dimensi spatial sebesar 224x224 untuk input I3D, dan 46 input koordinat xyz pada model ST-GCN.



Gambar 3.5. Metode perancangan model pada penelitian ini

Masing-masing model mengeluarkan output yang berisi matriks tensor yang mempunyai dimensi 1x50 dimana setiap elemen pada matriks merupakan skor klasifikasi pada masing-masing kata. Hasil skor klasifikasi pada masing-masing model kemudian akan dirata-ratakan menjadi klasifikasi akhir.

3.7 Pelatihan Model

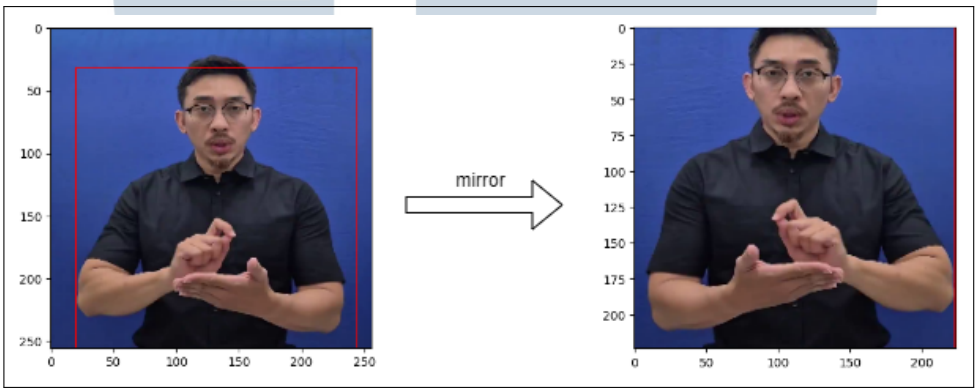
3.7.1 Augmentasi Data

Augmentasi video dilakukan untuk memperbanyak variasi sekaligus untuk memastikan setiap video mempunyai jumlah frame sebesar 64 dan dimensi sebesar 224x224 pada model I3D dan ST-GCN. Metode augmentasi video mengikuti praproses yang dilakukan pada penelitian sebelumnya. Augmentasi data meliputi augmentasi temporal (terhadap frame) dan augmentasi spatial (terhadap dimensi).

Untuk augmentasi temporal pada frame yang kurang dari 64, data video atau pose akan diberikan *padding* berupa frame awal atau frame akhir secara

acak dengan probabilitas 50%, untuk input aliran optik, akan diberikan *padding* berisi 0 dikarenakan tidak terdapat perubahan spatial pada video yang di *padding*. Sementara untuk ukuran frame yang lebih dari 64, akan diambil 64 frame subset yang berurutan pada sebuah video dengan indeks frame awal secara acak.

Sedangkan untuk augmentasi spatial akan dilakukan hanya pada model I3D dimana model akan mencari area 224x224 di video 256x256 secara acak yang kemudian akan dipotong pada seluruh frame seperti pada gambar 3.6. Setelah itu, video akan melakukan pembalikan horizontal (*mirroring*) secara acak dengan probabilitas 50% dikarenakan arti dari sebuah bahasa isyarat tidak akan berubah ketika di *mirror*.



Gambar 3.6. Metode augmentasi spatial video sebelum diberikan ke model

3.7.2 Parameter Tuning

Tuning parameter dilakukan dengan menggunakan parameter pada penelitian sebelumnya. Kedua model dilatih menggunakan *optimizer* Adam, dan dengan perhitungan nilai *loss* dilakukan menggunakan metode *categorical_crossentropy*. Nilai *epoch* pada kedua model sebesar 200 dan dilakukan *early stopping* dengan nilai *patience* sebesar 2 dengan memonitor nilai *val_accuracy*. Model I3D dilatih dengan parameter *learning rate* sebesar 10^{-3} dan *weight decay* sebesar 10^{-7}. Sedangkan model ST-GCN dilatih parameter *learning rate* sebesar 0.01 dan *weight decay* sebesar 10^{-4}. Model I3D dan ST-GCN dilatih menggunakan *batch_size* sebesar 6 dikarenakan keterbatasan memori pada perangkat keras selama penelitian.

3.8 Evaluasi Model

Evaluasi hasil model dilakukan dengan mengidentifikasi skor klasifikasi data testing dengan menggunakan metode top-N pada masing-masing kombinasi model dimana N = $\{1, 3, 5\}$. Penentuan nilai N dilakukan berdasarkan pada penelitian maruyama yang memakai nilai N = $\{1, 5, 10\}$, dikarenakan jumlah kata pada penelitian yang hanya mencapai 50 kata, sedangkan penelitian maruyama mempunyai 100 - 2000 kata. Maka di penelitian ini dilakukan penyesuaian rasio antara jumlah kata dengan banyaknya nilai N.

3.8.1 Kalkulasi Top-N

Tahapan pengubahan skor klasifikasi menjadi top-N dijelaskan sebagai berikut. Pertama-tama, prediksi akhir yang berupa skor klasifikasi berdimensi 1x50 akan diberikan label pada masing-masing kelas. Kemudian akan disortir berdasarkan skor klasifikasi dengan urutan skor klasifikasi terbesar sampai terkecil. top-1 adalah skor klasifikasi terbesar, sedangkan top-3 merupakan skor klasifikasi urutan pertama sampai ketiga (1-3), dan top-5 merupakan skor klasifikasi urutan pertama sampai kelima (1-5). Akurasi model akan di evaluasi dengan mengecek jika kelas dataset test masuk pada jangkauan prediksi top-N pada setiap data test.

3.8.2 Penentuan Kombinasi Model

Kombinasi model ditentukan dengan menggunakan kemungkinan seluruh kombinasi (misalnya rgb, rgb+farneback, rgb+farneback, skeleton, dst). Dikarenakan penelitian ini membandingkan dua tipe aliran optik (Farneback dan TV-L1), maka kombinasi model yang menggunakan aliran optik Farneback dan TV-L1 tidak digunakan (misalnya: rgb + tv-l1 + farneback). Sehingga kombinasi setiap model adalah sebagai berikut:

Tabel 3.2. Perancangan kombinasi model

No	model	rgb	farneback	TV-L1	skeleton
1	kombinasi1	✓			
2	kombinasi2		✓		
3	kombinasi3			✓	
4	kombinasi4				✓
5	kombinasi5	✓	✓		
6	kombinasi6	✓		✓	
7	kombinasi7	✓			✓
8	kombinasi8		✓		✓
9	kombinasi9			✓	✓
10	kombinasi10	✓	✓		✓
11	kombinasi11	✓		✓	✓

Kolom rgb merupakan model I3D yang menerima video instruktor dengan format rgb, sedangkan kolom farneback dan TV-L1 masing-masing merupakan model I3D yang menerima input hasil keluaran aliran optik Farneback dan TV-L1, sedangkan kolom *skeleton* merupakan model ST-GCN yang menerima input pose tubuh.