# BAB 2
# LANDASAN TEORI

## 2.1 Tinjauan Teori

Pada bagian ini akan diuraikan secara komprehensif dan mendalam teori-teori yang menjadi dasar penelitian. Tinjauan teori ini bertujuan untuk memberikan pemahaman yang lebih mendalam terhadap konsep-konsep yang relevan dan memperkuat dasar penelitian. Dalam tinjauan ini, akan dikaji berbagai sumber pustaka yang berkaitan dengan topik penelitian, termasuk jurnal ilmiah, buku, dan literatur primer maupun sekunder yang berhubungan dengan masalah yang diteliti. Pendekatan yang digunakan dalam penulisan tinjauan teori ini adalah pendekatan analitis dan sintesis untuk menyusun berbagai teori dan konsep yang relevan. Dengan demikian, tinjauan teori ini diharapkan dapat memberikan landasan yang kokoh bagi penelitian yang dilakukan.

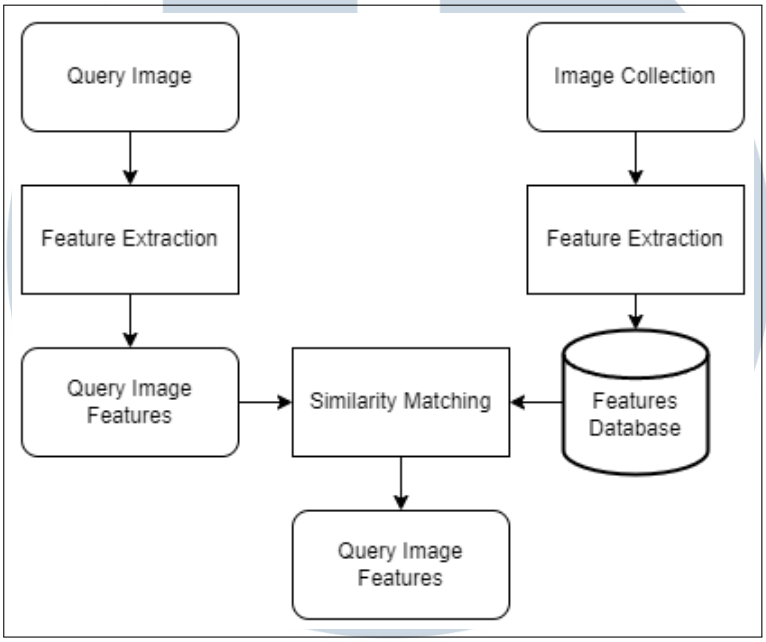
### 2.1.1 Text-Based Image Retrieval (TBIR)

*Text-Based Image Retrieval* (TBIR) adalah sebuah teknik yang digunakan untuk mencari gambar dengan menggunakan kata kunci, deskripsi, dan metadata lain yang terkait gambar tersebut. Metode ini memungkinkan pengguna untuk mencari gambar berdasarkan metadata yang disediakan. Namun, kekurangan dari penggunaan sistem TBIR adalah anotasi harus dimasukkan secara manual, sehingga dapat memakan waktu dan mungkin tidak sepenuhnya menangkap konten gambar. Selain itu, anotasi tekstual bergantung pada bahasa, yang dapat membatasi hasil pencarian[9].

### 2.1.2 Content-Based Image Retrieval (CBIR)

*Content-Based Image Retrieval* (CBIR) adalah teknik yang menggunakan visi komputer untuk mengambil gambar digital dari database berdasarkan konten visualnya. Hal ini berbeda dengan pencarian berbasis teks, yang menggunakan kata kunci dan keterangan untuk menemukan gambar. CBIR mencari gambar berdasarkan warna, tekstur, bentuk, dan fitur visual lainnya, dan mampu membedakan berbagai wilayah dalam sebuah gambar. Sistem CBIR dapat diklasifikasikan ke dalam dua jenis *query* yaitu berbasis teks dan berbasis gambar.

Dalam *query* berbasis teks, gambar ditentukan oleh informasi teks, seperti kata kunci dan keterangan. Sedangkan *query* berbasis gambar menggunakan contoh gambar untuk menemukan gambar yang serupa berdasarkan fitur-fiturnya, yang dapat diekstraksi secara otomatis[10].




Gambar 2.1. Arsitektur sistem CBIR

sumber: [10]

Sistem CBIR memiliki dua peran utama yaitu ekstraksi fitur dan pengukuran kemiripan. Ekstraksi fitur secara akurat mendefinisikan konten setiap gambar dalam database menggunakan fitur visual, dan informasi ini biasanya berukuran jauh lebih kecil daripada gambar aslinya. Pengukuran kemiripan melibatkan penghitungan jarak antara gambar *query* dan setiap gambar dalam database menggunakan fitur atau ciri-ciri yang telah diekstraksi, dan kemudian mengambil gambar ”terdekat” yang terbaik. Teknologi CBIR berguna untuk mengatur arsip gambar digital sesuai dengan konten visualnya dan menyediakan cara yang efisien untuk mencari dan mengambil gambar yang relevan[10].

### 2.1.3 Semantic-based Image Retrieval (SBIR)

*Semantic-based image retrieval* adalah teknik yang bertujuan untuk menemukan makna semantik tingkat tinggi di dalam sebuah gambar. Metode ini menggunakan informasi semantik untuk mengambil gambar dan termasuk

dalam kategori *Content-Based Image Retrieval*. Tidak seperti metode pencarian gambar tradisional, yang hanya berfokus pada fitur tingkat rendah, pencarian gambar berbasis semantik mencakup spektrum yang lebih luas dan bertujuan untuk menemukan gambar yang termasuk dalam kategori yang sama dengan *query*[5].

Tantangan utama dalam mewujudkan pengambilan gambar berbasis semantik adalah ”*semantic gap*”, yaitu perbedaan antara fitur tingkat rendah dan konsep tingkat tinggi. Untuk mengatasi kendala ini, pengambilan gambar berbasis semantik menggunakan pendekatan hybrid yang menggabungkan pendekatan berbasis bentuk, warna, dan tekstur untuk tujuan klasifikasi. Oleh karena itu, *Semantic-based image retrieval* adalah bidang penelitian yang aktif, dan para peneliti telah mengusulkan berbagai metode untuk meningkatkan akurasi pengambilan gambar yang relevan dengan *query* pengguna[11].

### 2.1.4 OpenAI CLIP Model

OpenAI *Contrastive Language-Image Pre-training* (CLIP) adalah jaringan saraf yang menghubungkan teks dan gambar. CLIP adalah model penglihatan dan bahasa multi-modal yang secara efisien mempelajari konsep visual dari pengawasan bahasa alami. CLIP diperkenalkan oleh OpenAI pada Januari 2021[12] dan dibangun di atas sejumlah besar penelitian tentang *zero-shot transfer*, *natural language supervision*, dan multimodal learning. Model ini dilatih pada berbagai pasangan (gambar, teks), di mana pasangan (gambar, teks) dapat berupa gambar dan keterangannya. CLIP dapat diinstruksikan dalam bahasa alami untuk memprediksi potongan teks yang paling relevan, jika diberikan sebuah gambar, tanpa secara langsung melakukan optimasi untuk tugas tersebut. Model ini menggunakan transformer mirip ViT untuk mendapatkan fitur visual dan model bahasa sebab akibat untuk mendapatkan fitur teks. Fitur teks dan visual kemudian diproyeksikan ke encoder penglihatan arsitektur CLIP. CLIP secara signifikan lebih fleksibel dan umum daripada model ImageNet yang ada dan dapat melakukan banyak tugas yang berbeda. CLIP dianggap sebagai salah satu kemajuan terpenting dalam visi komputer dalam beberapa tahun terakhir[6].

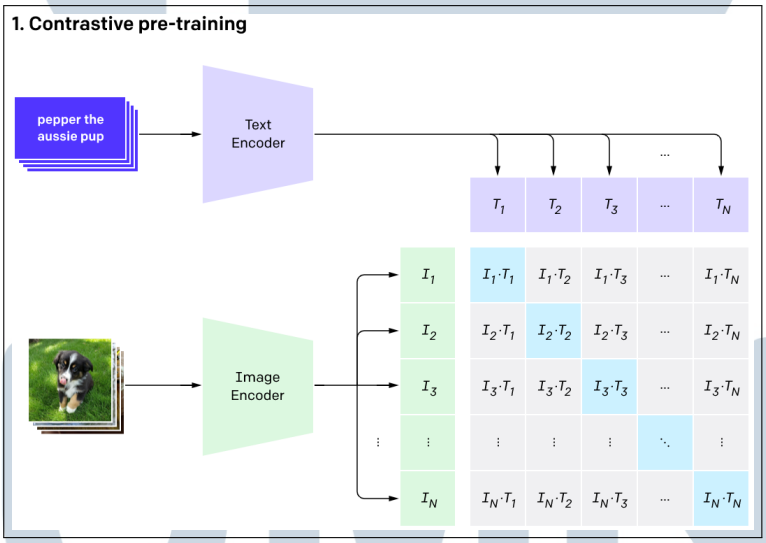
#### A CLIP Encoder Model

CLIP menggunakan model terpisah untuk *encoder* gambar dan *encoder* teks. *encoder* teks menggunakan transformer [13] dengan modifikasi arsitektur yang dijelaskan di [14]. Model ini menggunakan arsitektur Transformer dengan 12 *layer*, *width* 512, dan 8 *attention heads*. Untuk memproses masukan teks, teks mentah melalui proses enkoding pasangan byte [15], dengan *vocabulary size* sebesar 49.152. Panjang urutan teks dibatasi hingga 76 dan ditambahkan dengan

enkoding posisional sebelum dimasukkan ke dalam *encoder* teks.

Di sisi lain, CLIP menawarkan berbagai versi *encoder* gambar, termasuk arsitektur berbasis ResNet [16] dan Vision Transformer [17]. Mengingat kinerja yang superior yang ditunjukkan oleh model Vision Transformer dalam penelitian terkini [18], makalah ini secara eksklusif fokus pada *encoder* gambar berbasis Transformer. Seperti masukan teks, gambar-gambar dibagi menjadi beberapa *patch* dan setiap *patch* diberi enkoding posisional. Selanjutnya, dilakukan fungsi *global pooling* pada kedua *encoder*, yang mengompresi peta fitur menjadi satu fitur tunggal. Fitur ini berfungsi sebagai representasi dari seluruh gambar atau urutan teks, memungkinkan analisis dan pengolahan lebih lanjut.

## B CLIP Training




Gambar 2.2. Diagram pelatihan CLIP
sumber: [12]

Selama proses pelatihan, CLIP dilatih menggunakan pasangan gambar dan teks. Gambar-gambar tersebut awalnya diproses melalui model pemrosesan gambar, yang umumnya berdasarkan arsitektur resnet atau ViT. Langkah pengkodean ini mengubah gambar-gambar menjadi vektor *embedding* yang disebut ”I” dalam diagram. Secara bersamaan, teks yang terkait dengan setiap gambar diubah menjadi vektor *embedding* yang disebut ”T” menggunakan model transformer.

Penting untuk dicatat bahwa kedua *embedding* gambar dan teks memiliki

bentuk yang sama, memfasilitasi perbandingan langsung. Selama pelatihan, CLIP bertujuan untuk memaksimalkan kesamaan antara *embedding* gambar dan teks yang sesuai dengan pasangan gambar-teks yang sama, seperti yang ditunjukkan oleh diagonal biru dalam diagram. Sebaliknya, CLIP berusaha untuk meminimalkan kesamaan antara *embedding* yang terkait dengan pasangan yang tidak terkait.

Dengan mengoptimalkan proses ini, CLIP belajar untuk secara efektif mengaitkan dan membedakan antara *embedding* gambar dan teks yang relevan, memungkinkan pemahaman multimodal yang kuat dan kemampuan pengambilan kembali yang handal.

### 2.1.5 Cosine Similarity

*Cosine similarity* adalah sebuah perhitungan tingkat kemiripan antara dua vektor berdasarkan nilai kosinus dari sudut di antara keduanya. Ketika sudut kosinus adalah nol derajat, maka nilai yang dihasilkan adalah satu, tetapi untuk sudut lainnya, nilainya akan kurang dari satu dengan nilai minimum minus satu. Dengan demikian, nilai kosinus dari sudut antara dua vektor merupakan indikator apakah kedua vektor tersebut mengarah ke arah yang sama[19].

$$\cos(\theta) = \frac{A \cdot B}{\|A\| \, \|B\|} \tag{2.1}$$

Keterangan:

1. A : Vektor A
2. B : Vektor B
3. $A \cdot B$ : *Dot product* antara A dan B
4. $\|A\|$ : *Euclidian norm* dari Vektor A
5. $\|B\|$ : *Euclidian norm* dari Vektor B

Berikut adalah contoh yang menggambarkan perhitungan *cosine similarity*. Misalkan vektor A dan B didefinisikan sebagai berikut:

$$A = [5,6,9,7] \tag{2.2}$$

$$B = [0,5,2,7] \tag{2.3}$$

Untuk menghitung *cosine similarity* antara kedua vektor ini, dilakukan langkah-langkah berikut. Pertama, menentukan hasil perkalian *dot product* antara vektor-vektor tersebut:

$$A \cdot B = 5x0 + 6x5 + 9x2 + 7x7 = 97 \tag{2.4}$$

Selanjutnya, dihitung magnitudo dari vektor-vektor tersebut:

$$\|A\| = \sqrt{5^2 + 6^2 + 9^2 + 7^2} = \sqrt{191} \tag{2.5}$$

$$\|B\| = \sqrt{0^2 + 5^2 + 2^2 + 7^2} = \sqrt{78} \tag{2.6}$$

Terakhir, *cosine similarity* diperoleh dengan membagi hasil perkalian *dot product* dengan perkalian magnitudo:

$$\cos(\theta) = \frac{97}{\sqrt{191}\sqrt{78}} = 0,79 \tag{2.7}$$