Segementasi Optik Disc dan Cup untuk Membantu Pendeteksian Glaukoma Menggunakan *Segmentation Transformer*

1st Muhammad Raehan Akbar
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
mraehanakbar@student.telkomuniversity.ac.id

2nd Ema Rachmawati
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
emarachmawati@telkomuniversity.ac.id

3rd Mahmud Dwi Sulistiyo
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
mahmuddwis@telkomuniversity.ac.id

Abstrak-Glaukoma kondisi di mana saraf optik yang menghubungkan mata ke otak menjadi rusak. Glaukoma dapat menyebabkan kehilangan kemampuan penglihatan jika tidak didiagnosis dan ditangani secepat mungkin. Salah satu metode yang dilibatkan dalam mendiagnosis glaukoma menghitung rasio antara optik *disc* dan *cup* citra *fundus* mata. Untuk menghitung rasio antara *disc* dan *cup* citra *fundus* mata, diperlukan sebuah proses segmentasi citra *fundus* mata untuk dapat mensegmentasikan bagian *disc* dan *cup* nya. Saat ini tugas segmentasi dapat dilakukan menggunakan algoritma visi komputer modern. *Transformer* sendiri telah menjadi salah satu *state art of model* yang sering diterapkan studi kasus yang menggunakan *deep learning* karena performanya yang mampu menandingi *Convolutinal Neural Networks* (CNN). Tugas akhir ini akan membahas implementasi *Transformer* studi kasus segmentasi *disc* dan *cup* citra *fundus* mata menggunakan metode *Segmentation Transformer (SETR)* dengan dataset REFUGE dan DRISHTI-GS1. Hasil *dice coefficients score* dengan menggunakan *Cross Dataset Evaluation* berhasil mendapatkan skor 86 persen untuk bagian *disc* dan 78 persen untuk bagian *cup*.

Kata kunci - glaukoma, *disc*, *cup*, segmentasi, *segmentation transformers*, *transformers*.

Abstract-Glaucoma is a condition in which the optic nerve that connects the eye to the brain becomes damaged. Glaucoma can cause vision loss if not diagnosed and treated as soon as possible. One of the methods involved in diagnosing glaucoma is to calculate the ratio between the optic disc and cup on the fundus image of the eye. Calculating ratio between the disc and cup in the eye fundus image, a segmentation process is needed on the eye fundus image to be able segment the disc and cup parts. Nowdays the segmentation task can be performed using modern computer vision algorithms. Transformer itseld has become one of the state arts models that are often applied on case studies that use deep learning because the performance is proven to match the Convolutional Neural Netoworks (CNN). This final project will discuss the implementation of transformers in case study of disc and cup segmentation on eye fundus Images using the Segmentation Transformers (SETR) method with REFUGE and DRISHTI-GS1 datasets. The results of the dice coefficients score using Cross Dataset Evaluation managed to to get score 86 percent for the disc section and 78 percent for the cup section.

Keywords- glaucoma, disc, cup, segmentation, segmentation transformers, transformers.

I. PENDAHULUAN

A. Latar Belakang

Glaukoma merupakan kelainan mata yang disebabkan kerusakan saraf mata. Glaukoma penyebab kebutaan tertinggi ke 2 di dunia. Untuk mendiagnosa glaukoma, diperlukan tes yang dilakukan dokter mata salah satunya melakukan segmentasi bagian *disc* dan *cup* citra *fundus* keadaan mata yang dilakukan secara manual oleh dokter mata.

Masalah kemudian muncul segmentasi *disc* dan *cup* karena proses segmentasi secara manual citra *fundus* mata memakan banyak waktu oleh dokter mata karena bentuk *disc* dan *cup* yang harus dianotasi dan dihitung rasio antara *disc* dan *cup*. Untuk membantu dokter mata, dicetuskan terobosan untuk menyegmentasi optik *disc* dan *cup* secara otomatis dengan bantuan algoritma visi komputer untuk nantinya hasil segmentasi oleh algoritma visi komputer di analisa oleh dokter mata dengan menghitung rasionya untuk mendiagnosis glaukoma.

Penelitian terkait segmentasi *disc* dan *cup* banyak di kembangkan dimana diantaranya penelitian oleh Artem Sevastopolosky dkk [2] dengan menggunakan U-Net untuk menyegmentasikan bagian *disc* dan *cup* secara terpisah (tidak dalam satu citra *fundus* yang sama) dan penelitian yang dipaparkan oleh Syrna Sreng dkk [3] menggunakan DeepLabV3 menyegmentasikan bagian *disc* citra *fundus* mata.

Penelitian untuk menyegmentasikan *disc* dan *cup* mengalami perkembangan dalam tujuan dari segmentasi ,dengan menyegmentasikan *disc* dan *cup* citra *fundus* yang sama. Penelitian tersebut dilakukan Huanzhu Fu dkk [8] menggunakan U-NET.

tahun 2021 Shaohua Li dkk [4] berhasil mengimplementasikan dan mengembangkan model *transformer* untuk studi kasus segementasi gambar

medis dimana studi kasus yang di selesaikan segmentasi *disc* dan *cup* dataset REFUGE [5] satu citra *fundus* yang sama dengan model yang dinamai SEGTRAN. Selain SEGTRAN, Shaouha Li dkk [4] turut melakukan uji coba implementasi menggunakan model *transformer* untuk segementasi yang dikembangkan oleh Sixiao Zheng [1] yaitu SETR (Segmentation Transformer) dengan mencoba salah satu bentuk *decoder* SETR. Tugas akhir ini akan menguji SETR dengan variasi bentuk bentuk *decoder* studi kasus segmentasi *disc* dan *cup* *citra fundus* mata.

B. Identifikasi Masalah

Masalah yang teridentifikasi tugas akhir ini bagaimana membangun sebuah system segmentasi bagian *disc* dan *cup* satu citra *fundus* dengan menerapkan Teknik *Computer Vision* dan *Deep Learning?*.

C. Batasan Masalah

Adapun Batasan masalah tugas akhir ini sebagai berikut

1. Dataset yang digunakan dataset REFUGE [5] dan DRISHTI-GS1 [6].
2. Bentuk *Decoder* SETR yang di implementasikan tugas akhir ini telah mengalami *pre-trained* menggunakan dataset *imagenet21k* [12].
3. Dataset DRISHTI-GS1 [6] *cropping* bagian lokasi *disc* dan *cup* dilakukan manual.

D. Tujuan

Tujuan tugas akhir ini mengimplementasikan dan mengevaluasi SETR kasus segementasi *disc* dan *cup* studi kasus untuk membantu pendeteksian glaukoma.

E. Kegiatan Penelitian

1. Kajian Pustaka
   proses ini kegiatan yang dilakukan mempelajari teori yang berkaitan dengan tugas akhir ini seperti mempelajari dataset yang berkaitan dengan segmentasi *disc* dan *cup,* metode yang dikembangkan untuk melakukan segmentasi *disc* dan *cup,* dan implementasi *Transformers* berbagai studi kasus khususnya segmentasi.
2. Pengumpulan Data
   proses ini kegiatan yang dilakukan mengumpulkan dataset yang digunakan dalam penelitian tugas akhir ini.
3. Perancangan Sistem
   proses ini kegiatan yang dilakukan merancang sistem yang dibangun untuk penelitian tugas akhir dengan menggunakan model yang diajukan tugas akhir.
4. Pengujian Tugas Akhir
   proses ini kegiatan yang dilakukan melakukan pengujian terhadap model yang sebelumnya dirancang sistem menggunakan dataset yang dikumpulkan.
5. Analisis Hasil dan Penulisan Laporan Tugas Akhir
   proses ini kegiatan yang dilakukan melakukan analisis hasil pengujian tugas akhir dan di tuliskan ke dalam bentuk laporan untuk mendapatkan kesimpulan akhir dan saran.

II. KAJIAN TEORI

A. Segmentasi Disc dan Cup Untuk Mendeteksi Glaukoma

Dataset REFUGE [5] dataset yang disediakan kompetisi *Retinal Fundus Glaucoma Challenge* berkonjungsi dengan MICCAI (*Medical Imaging and Computer Assisted Invervention conference)* 2020. Kompetisi ini menantang pesertanya menyelesaikan dua tantangan wajib yaitu mengklasifikasikan glaukoma dan non-glaukoma dan menyegmentasikan optik *disc* dan *cup.*Satu tantangan bonus menentukan lokasi *fovea* citra *fundus* mata yang disediakan pihak penyelenggara. Kompetisi ini untuk mengembangkan penelitian bidang deteksi glaukoma secara otomatis menggunakan *Deep Learning*.

Serupa dengan REFUGE, dataset DRISHTI-GS1 disediakan pihak CVIT *(Center for Visual Information Technology)* untuk penelitian deteksi glaukoma. Dataset ini berasal dari kompetisi yang diadakan pihak CVIT, yang membedakannya bentuk *masking* yang disediakan DRISHTI-GS1anotasi terpisah *disc* dan *cup.* Sedangkan REFUGE bentuk anotasinya dalam satu gambar *(masks* bentuk *disc* dan *cup).*

Salah satu penelitian yang menggunakan dataset REFUGE, penelitian yang dikembangkan Shaoua li dkk [4] menggunakan *transformer*, menambahkan *Squueze* dan *Expansion* blok *Attention* layer *transformer* dengan sebutan SEGTRAN. Selain memaparkan SEGTRAN-nya, Shaouha li dkk [4] memaparkan hasil SEGTRAN dengan model segmentasi pendekatan *transformer* lainnya salah satunya SETR yaitu model yang dikembangkan oleh Shiaou Zeng dkk [1]. Skor rata-rata *Dice Coefficient* terhadap *disc* dan *cup* yang didapatkan SEGTRAN maksimal 91.7 % sedangkan SETR 90.5 %. penelitiannya Shaohua li dkk [4] menggunakan DRISHTI-GS1 sebagai dataset tambahan yang digunakan untuk data latih penelitiannya.

B. Transformer

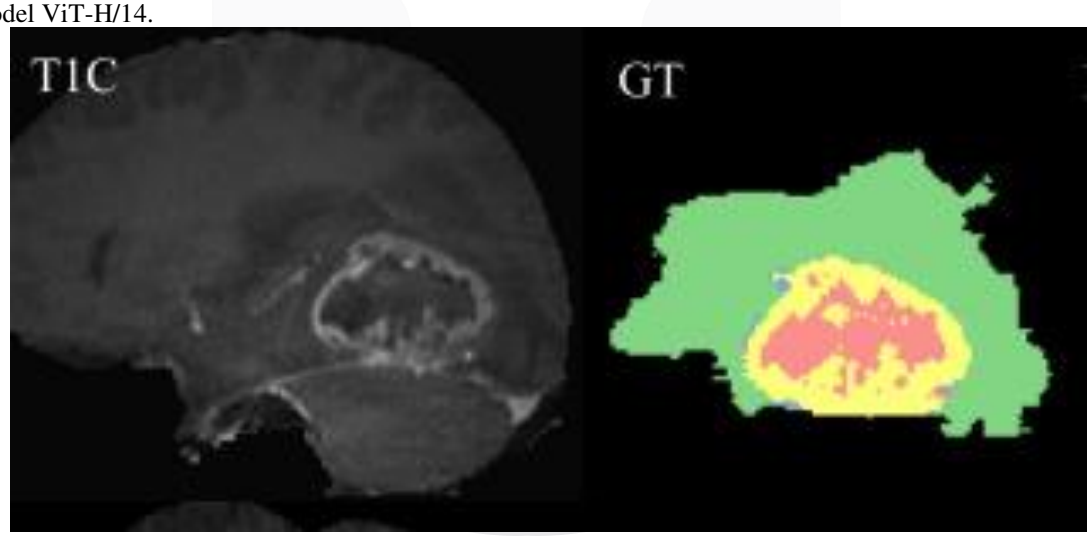
Transformer pertama diperkenalkan tahun 2017 oleh Ashish Vaswani dkk [7]. Model Arsitektur ini memiliki penghubungan antara *encoder* dan *decoder* dengan mekanisme *attention yang murni* tanpa *recurrence* dan *convolution. Transformer* dikembangkan sebagai alternatif kebanyakan model arsitektur yang menggunakan *recurrence* dan *convolutional* yang kompleks. Model arsitektur diuji tantangan untuk menerjemahkan bahasa Inggris ke bahasa Jerman dan mendapatkan skor *BLEU* 28.4 lebih baik dari model *ConvS2S Ensemble* yang Arsitektur model ini menggunakan *convolution* sebagai konsep dasarnya dan mendapat skor *BLEU* 26.36 .

Setelah kesuksesan *transformer* Pemrosesan bahasa alami, penelitian ini dikembangkan untuk diterapkan visi komputer. Kemudian di tahun 2020 Alexey Dosovitsky dkk [10] memperkenalkan metode yang bernama *Vision Transformer* (ViT) model arsitektur visi komputer yang menggunakan *transformer* implementasinya yang digunakan untuk mengklasifikasikan objek citra. Model ini telah di ujikan dataset ImageNet [12] untuk menglasifikasikan objek yang beragam citra dengan akurasi terbaik 88.55 dengan variasi model ViT-H/14.

Setelah keberhasilan ViT, perkemabangan *transformer* visi komputer semakin pesat. Hasilnya tahun yang sama ditemukannya ViT, *Detection Transformer* (DETR) ditemukan Nicolas Carion dkk [9] yang digunakan untuk mendeteki objek citra menggunakan *transformer* sebagai implementasinya. Terbaru beberapa pengembangan *transformer* untuk melakukan segmentasi citra pun dikembangkan seperti SEGTRAN oleh Shaouha li dkk [4] dan SETR oleh Sixiao Zheng dkk [1].

C. Medical Image Segmentation

Medical Image Segmentation bertujuan menyegmentasi objek *medical image* seperti MRI,CT dan X-RAY untuk mendeteksi masalah kesehatan seperti tumor, kerusakan suatu jaringan syaraf , dan masalah kesehatan lainnya. Tugas menyegmentasi memakan banyak waktu karena dilakukan secara manual untuk melakukan segmentasi bagian bentuk citra tersebut.Modern ini tugas tersebut dibantu menggunakan *Deep Learning* untuk mempermudah segmentasi sehingga dokter fokus untuk melakukan diagnosa berdasarkan hasil dari *Deep Learning*.



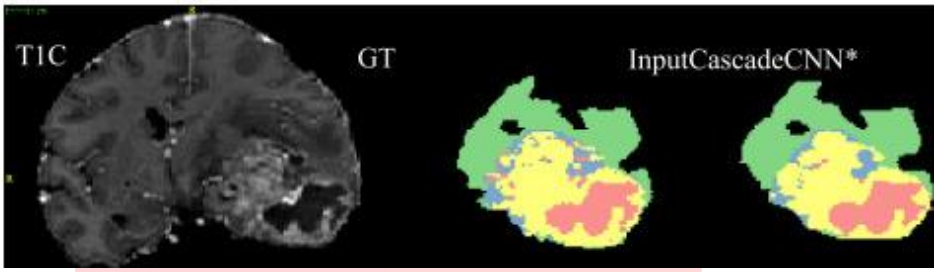
GAMBAR 1:
CONTOH *MEDICAL IMAGE SEGMENTATION* TUMOR OTAK [11]

Ilustrasi diatas contoh *Medical Image Segmentation* studi kasus segmentasi tumor otak citra MRI otak. Terlihat terdapat beberapa bagian yang teridentifikasi sebagai bagian dengan tumor otak/. Implementasinya *Medical Image Segmentation* menerapkan *Semantic Segmentation* yaitu segmentasi melabeli sebuah obyek per piksel citra berikut tahapannya.

1. Citra asli dilakukan anotasi dengan metode *masking* yaitu pewarnaan piksel yang ingin diberi label. Seperti ilustrasi diatas
   a. Busung (Edena) warna hijau.
   b. Tumor yang berkembang (*Advancing tumor)* warna kuning
   c. Tumor yang tidak berkembang (*Non-Advancing tumor)* warna biru.
   d. Inti tumor nekrotik (*Necrotic tumor core)* warna merah
2. Setelah *masking* citra selesai, data dipilah menjadi 2 bagian yaitu data citra

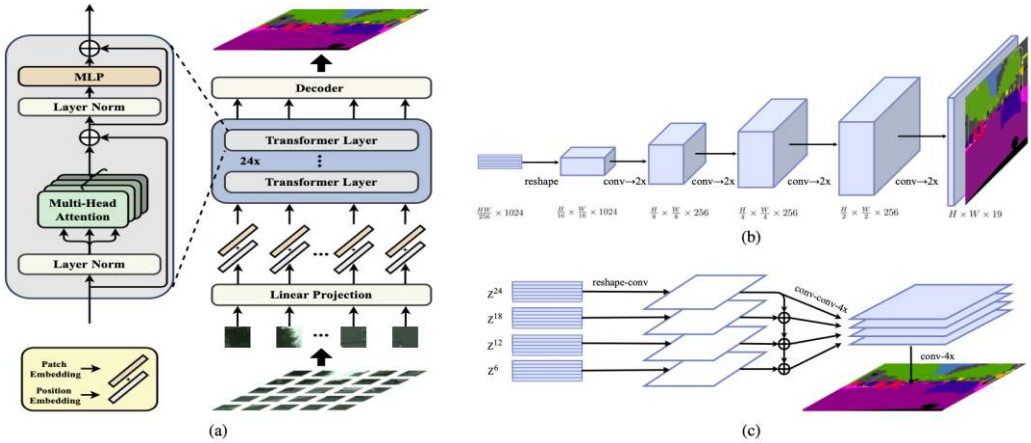
asli dengan citra yang telah di *masking* sebagai *ground* truth.

3. Terakhir aproses pelatihan data model. Metode yang digunakan menggunakan teknik *Convolutional Neural Networks*. Ssederhananya citra asli diinputkan model *convolutional Neural Network* kemudian prosesnya mengalami *convolution pooling,* dan *upsampling* untuk mendapatkan keluaran label piksel dengan kelas nya. Keluaran dari piksel akan dibandingkan dengan *ground truth* nantinya model *Convolutional Neural Network* dapat mempelajari *ground truth* untuk memberikan keluaran segmentasi yang lebih baik.



GAMBAR 2:
PERBANDINGAN INPUT GAMBAR, GROUND TRUTH, DAN HASIL PREDIKSI SEGMENTASI GAMBAR [11]

D. Segmentation Transformers (SETR)



GAMBAR 3:
(A) ARSITEKTUR SETR, (B) DESAIN *DECODER* PUP, (C) DESAIN DECODER MLA [1]

Segmentation Transformers (SETR) terdapat 3 inti yaitu *Image to Sequence, Transformers Layer,* dan *Decoder Design.* Model ini memperkenalkan segmentasi semantik dengan bentuk *sequence to sequence.* Sederhananya model ini mengubah citra ke bentuk *sequence* satu dimensi kemudian diolah *layer* transformer. *Transformer* berperan sebagai *encoder* menjadi paramater latih.Keluaran dari *transformer* akan diolah *decoder* untuk segmentasi *pixel-level* yang menghasilkan sebuah prediksi segmentasi. Berikut penjelasannya.

a. Image To Sequence

tahap ini citra masukan diubah ke bentuk 1D *sequences.* karena bentuk masukan yang diterima oleh SETR $Z \in \mathbb{R}^{L \times C}$ dengan L panjang *sequence* dan C ukuran *channel* sedangkan citra memiliki bentuk $x \in \mathbb{R}^{H \times W \times 3}$, maka *input* citra berbentuk $x \in \mathbb{R}^{H \times W \times 3}$ perlu diubah kebentuk Z. Karena bentuk *encoder semantic segmentation* men-*downsample* citra 2D berbentuk $x \in \mathbb{R}^{H \times W \times 3}$ ke bentuk *feature map* $x_f \in \mathbb{R}^{\left(\frac{H}{16}\right) \times \left(\frac{W}{16}\right) \times C}$ maka L bentuk yang diterima oleh SETR dapat dirubah ke bentuk $\left(\frac{H}{16}\right) \times \left(\frac{W}{16}\right) = \frac{HW}{256}$ dengan demikian *output* dari *sequence trasnformers* dapat di ubah ke bentuk *feature map* x_f. untuk mendapatkan *sequences* sepanjang $\frac{HW}{256}$ citra berbentuk $x \in \mathbb{R}^{H \times W \times 3}$ di bagi ke bentuk grid $\left(\frac{H}{16}\right) \times \left(\frac{W}{16}\right)$ secara *uniform*, grid di *flatten* ke bentuk *sequence.* Kemudian tiap patch p yang divektorisasi di *mapping* menuju C-*dimensional embedding space* dengan fungsi

projeksi linear $f: p \rightarrow e \in \mathbb{R}^C$. Dari fungsi tersebut *sequence* satu dimensi hasil dari *patch embeddings* untuk citra x, di dapatkan dengan bentuk $E = \{e_1 + p_1, e_2 + p_2, \ldots, e_L + p_L\}$.

b. Transformers Layers

Setelah *sequence* 1D dengan bentuk *E* dijadikan *input encoder*, *Transformer encoder* digunakan untuk mempelajari *feature representations.* Setiap *layers transformer* memiliki *global receptive field.* Setiap *encoder transformer* berisikan L_e *layers* dengan blok MSA(*multi-head self-attention*) dan blok MLP(*multi-layer-perceptron*). setiap *layer l input* menuju *self-attention* yang berada dalam triplet (*query, key, value*) dihitung dari input $Z^{(l-1)} \in \mathbb{R}^{L \times C}$ sebagai:

$$query = Z^{(l-1)}W_Q, key = Z^{(l-1)}W_K, value = Z^{(l-1)}W_V. \quad (1)$$

Di mana $W_Q/W_K/W_V \in \mathbb{R}^{C \times d}$ parameter latih 3 *layer* proyeksi linier dan *d* dimensi (*query, key, value*). Kemudian *Self-Attention* (SA) di formulasikan:

$$SA(z^{l-1}) = z^{l-1} + softmax\left(\frac{(z^{l-1}W_Q(ZW_K)^T)}{\sqrt{d}}\right)(Z^{l-1}W_V). \quad (2)$$

Untuk MSA ekstensi dari operasi independen SA sebanyak *m* dan keluarannya di gabung menjadi :

$$MSA(Z^{l-1}) = [SA_1(Z^{l-1}); SA_2(Z^{l-1}); \ldots; SA_m(Z^{l-1})]W_O$$

. dimana $W_O \in \mathbb{R}^{md \times C}$ nilai *d* di set menjadi *C/m*. Keluaran dari MSA di transformasikan blok MLP bersamaan *residual skip* sebagai layer *output* sebagai

$$Z^l = MSA(Z^{l-1}) + MLP\big(MSA(Z^{l-1})\big) \in \mathbb{R}^{L \times C}. \quad (3)$$

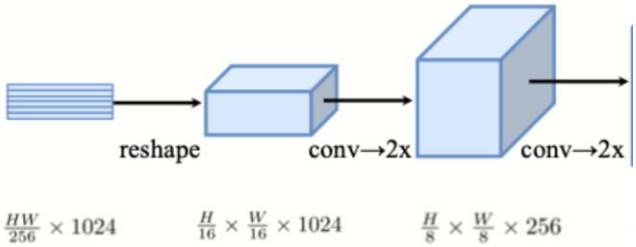
Dengan *layer norm* di aplikasikan sebelum blok MSA dan MLP persamaan (3) tidak ditampilkan untuk meringkas. Dengan demikian layer *transformers* $Z^1, Z^2, \ldots, Z^{Le}$.

c. Decoder Design

Setelah *sequence* 1D berbentuk *E* dijadikan *input encoder*, diperlukan *decoder* untuk *pixel-level segmentation* untuk mendapatkan gambar segmentasi setiap piksel nya. Sebagaimana *decoder* diperlukan untuk menghasilkan citra 2D ($H \times W$), maka feature *encoder* dengan reprentasi Z perlu di *reshape* dari 2D $\left(\frac{HW}{256}\right) \times C$ ke standar *feature map* 3D $\left(\frac{H}{16}\right) \times \left(\frac{W}{16}\right) \times C$ SETR memiliki 3 bentuk *decoder* yaitu *Naive upsampling, Progressive Upsampling(PUP), Multi -Level feature Aggregation (MLA).* Berikut penjelasan ketiga bentuk decodernya.

a) Naive Upsampling (Naive)

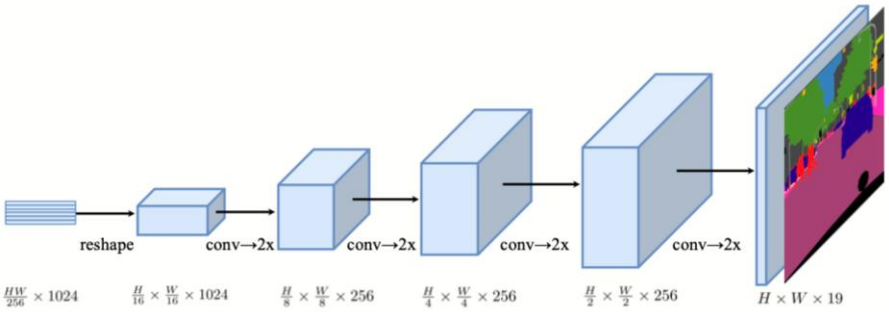
Naive *decoder* memproyeksikan *transformer feature* Z^{Le} ke dimensi jumlah kategori. Untuk proyeksi tersebut digunakan *simple layer network* yang berjumlah 2 *layer* dengan arsitektur: 1×1 *conv* + *sync batch norm (w/ReLU)* + 1×1. Kemudian *output*-nya di *upsample* secara biliner ke bentuk citra resolusi penuh.



GAMBAR 4 :
ILUSTRASI DESAIN *DECODER NAÏVE.*

b) Progressive Upsampling (PUP)

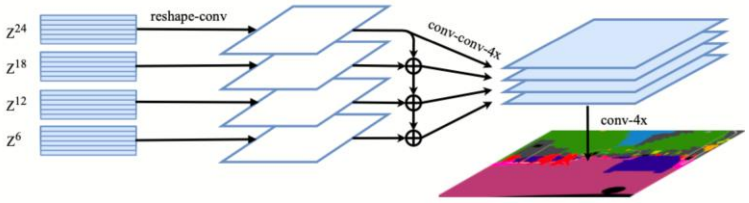
Dengan pretimbangan *Naive upsampling* mungkin menghasilkan prediksi yang kurang baik, untuk memaksimalkan efek *adversarial,* upsampling ditambahkan hingga 2 kali. Sederhananya jika *Naive upsampling* menggunakan 2 *layer* arsitektur maka PUP menggunakan 4 layer dengan arsitektur untuk mencapai resolusi penuh dari Z^{L_e} dengan ukuran gambar $\left(\frac{H}{16}\right) \times \left(\frac{W}{16}\right)$.



GAMBAR 5:
ILUSTRASI DESAIN *DECODER PUP*

c) Multi-Level Feature Aggregation (MLA)

Bentuk *decoder* mengambil *feature representations* $\{Z^m\}$ $\left(\left(m \in \left\{\frac{L_e}{M}, 2\frac{L_e}{M}, \ldots, M\frac{L_e}{M}\right\}\right)\right)$ dari *layers M* yang secara *uniform* terdistribusi sepanjang *layers* dengan *step* $\frac{L_e}{M}$ decoder. Kemudian *streams* sebanyak *M* dikerahkan dengan setiap *streams* di fokuskan spesifik satu *layer* pilihan. tiap *stream, feature encoder* Z^l dari bentuk 2D $\frac{HW}{256} \times C$ di *reshape* kebentuk 3D *feature map* $\frac{H}{16} \times \frac{W}{16} \times C$. Untuk melakukannya digunakanlah 3-*layer network(kernel size* 1×1, 3×3, dan 3×3*)* dengan *feature channel* layer pertama dan terakhir dibelah dan *spatial resolution* di *upscaled* hinggan 4 kali dengan operasi bilinier setelah *layer* terakhir. Untuk mendapatkan interaksi sepanjang *streams,* digunakanlah desain *top-down aggregation* melalui penambahan *element-wise* setelah *layer* pertama. Kemudian *layer convolutional* 3×3 tambahan di terapkan setelah *element-wise* melakukan penambahan *feature.* Setelah l*ayer* terakhir. Semua fitur gabungan yang didpatkan dari semua *streams* melalui penyatuan channel-wise kemudian di *upsample* secara bilinier sebanyak 4 kali untuk mendapatkan resolusi penuh.



GAMBAR 6 :
ILUSTRASI DESAIN *DECODER* MLA

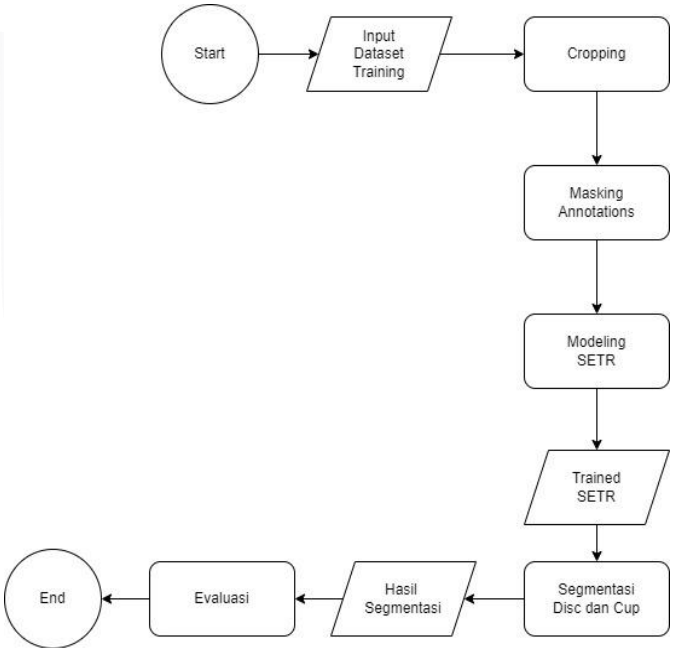
Untuk jenis *backbone*(variasi bentuk encoder transformer) yang disebut sebagai backbone yaitu *T-Base* dan *T-Large*. Perbedaan *T-Base* dan *T-Large* dapat dilihat tabel 1.

TABEL 1:
VARIASI *BACKBONE* SETR

Model Backbone	Jumlah Layer Transformer	Jumlah Hidden Size Transformer	Jumlah Head Attention transformer
T-Base	12	768	12
T-Large	24	1024	16

III. METODE

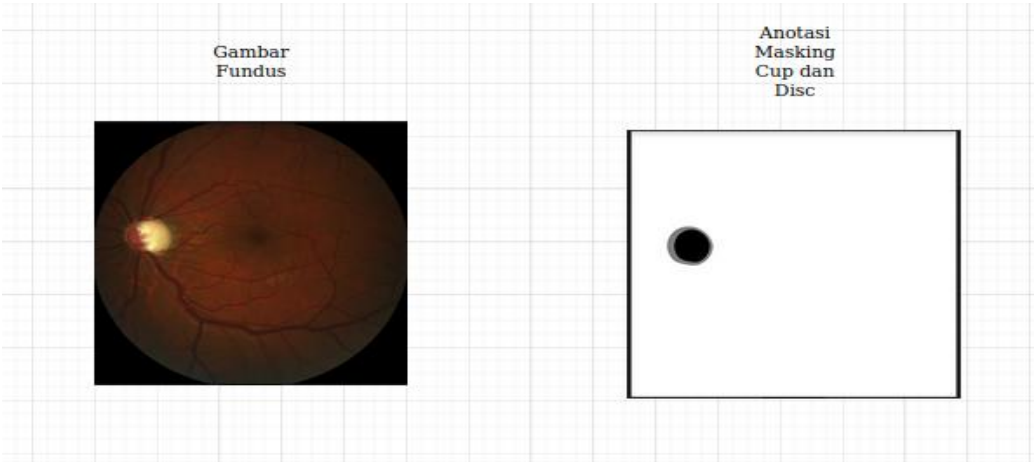
Penelitian ini menggunakan model *Segmentation Transformer* (SETR) untuk mensegmentasi bagian *disc* dan *cup* citra *fundus.* Keluaran dari model ini sebuah citra 2D dengan segmentasi 2 warna yaitu abu abu (bagian *disc*), dan hitam (bagian *cup*).



GAMBAR 7:
ALUR SISTEM SEGMENTASI *DISC* DAN *CUP* DENGAN SETR

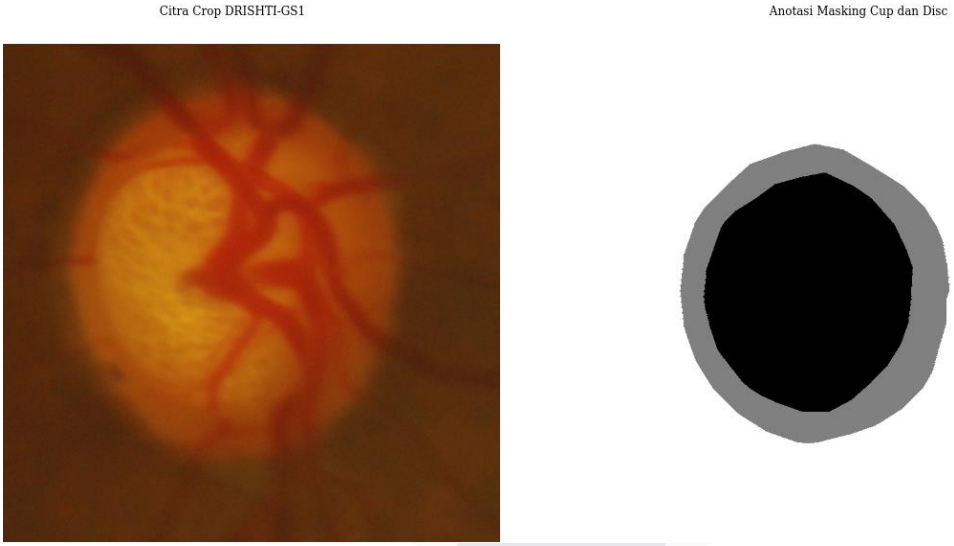
A. Dataset

Dataset yang digunakan berasal dari 2 sumber yaitu REFUGE [5] dan DRISHTI-GS1 [6]. Dataset REFUGE merupakan dataset yang disediakan kompetisi REFUGE 2020 [5]. Dataset ini terdiri atas 3 *folder* yaitu *Training, Validation,* dan *Test* dengan setiap *folder* memiliki citra asli berupa citra *fundus* keadaan mata yang bertipe data *jpg* dan anotasi berupa *masks disc* dan *cup* yang bertipe data *bmp* dan *masks* tersebut berada satu citra yang sama. Jumlah citra REFUGE [5] secara total 1200 citra dengan rincian 400 citra untuk *Training,* 400 citra untuk *Validation,* dan 400 citra untuk *Test.*



GAMBAR 8:
CITRA *FUNDUS* ASLI DAN ANOTASI *MASKS DISC* DAN *CUP* DATASET REFUGE

Dataset DRISHTI-GS1 merupakan dataset yang disediakan untuk penelitian dibidang glaukoma oleh CVIT. Dataset ini terdiri atas 2 *folder* yaitu *Training* dan *Test* dimana tiap folder memiliki citra asli *fundus* mata dengan *masks disc* dan *cup*. Berbeda dengan REFUGE, DRISHTI-GS1 memliki *masks disc* dan *cup* yang terpisah oleh karenanya dilakukan *image processing* oleh Shaoua Li pengembang SEGTRAN untuk mendapatkan bentuk *masks* yang serupa dengan REFUGE yaitu satu citra yang sama. Jumlah citra DRISHTI-GS1 secara total 101 citra dengan rincian *Training* 50 citra dan *Test* 51 Citra.



GAMBAR 9:
CITRA *FUNDUS* ASLI DAN ANOTASI *MASKS DISC* DAN *CUP*
DATASET DRISTHI-GS1 SETELAH *IMAGE PROCESSING*

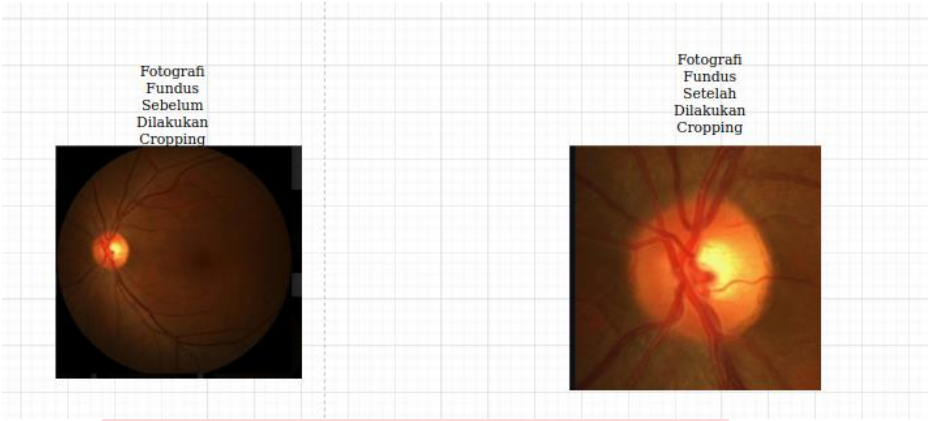
B. Preprocessing

tahapan ini dilakukan *preprocessing* citra yang melibatkan 2 proses yaitu *Cropping* citra *fundus* dan *Masking Annotation.*

1. Cropping Citra Fundus

Tahapan ini tahapan untuk melakukan cropping citra utuh fundus menjadi citra yang hanya menampilkan bagian yang di identifikasi sebagai bagian disc dan cup atau disebut juga sebagai ROI(Region of Interest). Tahapan ini pun menggunakan cropping yang manual dan otomatis. Untuk cropping ROI secara manual dilakukan dataset DRISHTI-GS1 menggunakan photo editor GIMP. Kemudian setelahnya hasil cropping tersebut di resize menggunakan fungsi untuk melakukan resize citra fundus MNET DEEP CDR [8] dengan ukuran 576×576.

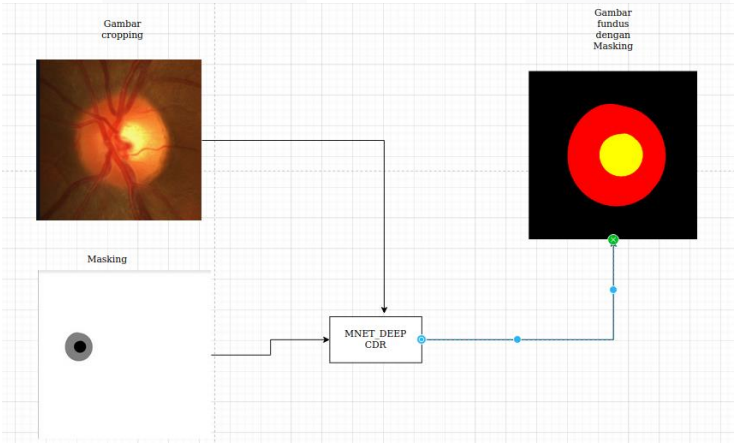
Untuk cropping secara otomatis REFUGE turut menggunakan MNET DEEP CDR dengan fungsi untuk melakukan deteksi secara otomatis bagian ROI citra fundus REFUGE untuk kemudian di cropping dengan ukuran 576×k dengan k bergantung ukuran hasil deteksi MNET DEEP CDR.



GAMBAR 10:
PERBANDINGAN SALAH SATU CITRA DATASET REFUGE SEBELUM DI CROPPING (KIRI) DAN SESUDAH (KANAN) UNTUK MENDAPATKAN ROI

2. Masking Annotation

tahapan ini citra fundus yang telah di cropping, di anotasi bagian yang di identifikasi sebagai disc dan cup dengan metode yang masking. Singkatnya anotasi ini dilakukan dengan cara mewarnai bagian yang menjadi ROI yaitu warna merah untuk bagian disc dan kuning untuk bagian cup berdasarkan bentuk masks bawaan. Meski tiap dataset telah memiliki masks, namun anotasi ini diperlukan karena setelah cropping bentuk masks harus disesuaikan dengan kondisi masks yang menggunakan citra asli. Proses masking annotation dibantu dengan framework MNET DEEP CDR.



GAMBAR 11:
PROSES *MASKING ANNOTATION* DATASET

C. Pemodelan Menggunakan SETR

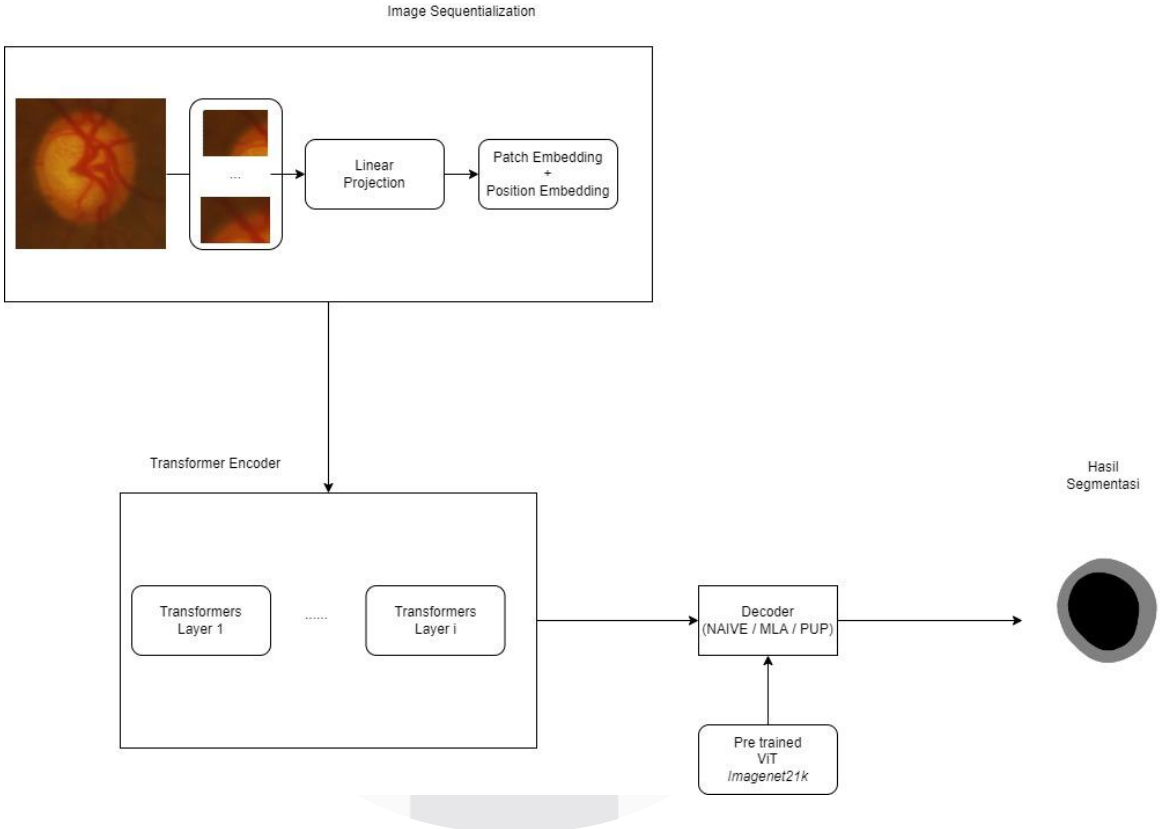
tahapan ini pemodelan dilakukan dengan implementasi menggunakan SETR. Implementasi penggunaan SETR penelitian ini menggunakan repository oleh Shoauha Li dan koleganya yaitu SEGTRAN yang mana repository tersebut telah melakukan implmentasi penggunaan SETR dengan salah bentuk decoder SETR PUP. Sebagai improvisasi dari repository tersebut, penggunaan SETR pun ditambahkan dengan mengimplementasikan 2 bentuk decoder lainnya dari SETR yaitu MLA dan NAIVE. *Transformer layer* SETR ini sebelumnya telah dilakukan pre-trained weights *transformer layer* yang digunakan Vision Transformer (ViT) [10] dengan dataset

Imagenet21K [12]. Berikut tabel informasi SETR yang digunakan penelitian ini.

TABEL 2:
BENTUK *DECODER* DAN INFORMASI JUMLAH PARAMETER SETIAP BENTUK *DECODER* YANG DIGUNAKAN

Bentuk Decoder	Backbone	Jumlah Parameter
SETR-Naive	T-Large	305.67 Juta
SETR-MLA	T-Large	310.57 Juta
SETR-PUP	T-Large	318.31 Juta

Untuk tahapan proses segmentasi, seperti yang telah dijelaskan bab 2.4 dimana gambar citra asli dari dataset akan dipecah menjadi beberapa ukuran *patches* untuk, diproses *linear projection* untuk *embed* tiap *patch* tersebut, ditambahkan *position embeddings,* dimasukan *encoder transformer,* dan mendapat *pixel-wise segmentation decoder* untuk kemudian mendapat keluaran berupa hasil prediksi *mask* citra yang menjadi masukan. Penjelasan singkat proses SETR dapat dilihat gambar berikut.



GAMBAR 12 :
PROSES SEGMENTASI *DISC* DAN *CUP* DAN SETR

D. Metrik Pengukuran

Untuk menguji kinerja SETR, metrik pengukuran yang digunakan penlitian ini ada 2 yaitu *Dice-coefficient score* dan *Average CDR Error.*

1. Dice-coefficient Score

Dice-coefficient Score atau disebut juga *Dice Score* metrik pengukuran yang digunakan untuk mengukur similaritas 2 buah citra. Metrik ini digunakan untuk mengukur hasil dari keluaran segmentasi oleh SETR dan ground-truth setelah proses masking nya. Namun penelitian yang akan ditampilkan rata rata dari *Dice Score* Implementasi *Dice-coefficient* yang digunakan penelitian ini.

$$overlap\ disc\ = (2 * area\ over\ lap\ disc) \ / \ (total\ piksel)$$
$$overlap\ cup\ = (2 * area\ over\ lap\ cup) \ / \ (total\ piksel\ )$$

$$overlap \text{ BG } = (2 * area\ overlap \text{ BG } / total\ piksel)$$
$$disc\ dice\ score\ (D_d) = (overlap\ disc\ + overlap\ BG)/2$$
$$cup\ dice\ score\ (D_c) = (overlap\ cup\ + overlap\ BG)/2$$

BG = Backgorund

2. Average CDR Error (vCDR Error)

Persamaan ini untuk menghitung rerata dari error CDR(Cup to Disc Ratio) dari masks hasil prediksi dan masks data uji(masks sebelum cropping). Nantinya nilai error tiap citra ini dijumlahkan untuk di bagi dengan jumlah citra data uji. Untuk penghitungan error tiap citra dirumuskan dengan formula berikut

$$E_{vcdr}(S,G) = 1 - \frac{(Area(S \cap G))}{(Area(S \cup G))} \quad (4)$$

dengan S area hasil segmentasi SETR dan G area mask Ground Truth

sebagai contoh untuk menghitung vCDR Error data uji DRISHTI-GS1 dengan dataset test berjumlah n

$$E_{vcdr}\ Drishti = \frac{E_1 + E_2 + \ldots + E_n}{n} \quad (5)$$

E. Fungsi Loss

Fungsi loss penelitian ini *Dice loss* dan *Cross Entropy*

1. Dice Loss

Dice loss fungsi loss yang digunakan untuk menghitung similaritas antar dua citra. Fungsi ini merupakan penghitungan loss yang berasal dari Dice Score. Berikut persamaannya

$$D_l(p_i, g_i) = 1 - \frac{2p_i g_i + 1}{p_i + g_i + 1} \quad (6)$$

dengan p_i jumlah piksel hasil prediksi dan g_i merepresentasikan jumlah piksel *ground truth.* Penambahan 1 berfungsi sebagai *numerator* dan *denominator* unutk memastikan fungsi tidak terdefinisi seperti $p_i = g_i = 0$

.

2. Cross Entropy

Cross Entropy fungsi *loss* yang umum digunakan tugas image segmentation *Cross Entropy* yang digunakan *pixel-wise cross entropy*. Fungsi menghitung perbedaan antar 2 probabilitas distribusi. Fungsi ini digunakan digunakan untuk menghitung seberapa dekat suatu distribusi kemunculan acak dengan kemunculan acak lainnya. Dengan demikian fungsi *loss* ini digunakan untuk perbedaan konten informasi antara hasil prediksi dan *ground truth*. Berikut persamaannya.

$$\text{CE}(g_i, p_i) = -(g_i \log(p_i) + (1 - g_i) \log(1 - p_i)) \quad (7)$$

Dengan demikian *pixel-wise loss* dihitung dengan *log loss* dan dijumlahkan dengan setiap kemungkinan kelas. Penghitungan kemudian diulang setiap piksel dan di rata-rata kan.

IV. HASIL DAN PEMBAHASAN

A. Hasil Pengujian

Pengujian peneltian ini dilakukan dengan 4 skenario yang berbeda. *Hyperparameter* yang digunakan untuk ke 4 skenario menggunakan Adam sebagai *optimizer*, nilai *learning rate* 0.0002, nilai *batch size* 6, nilai *decay* 0.0001, ukuran *patch size* (penyesuaian ukuran citra sebelum dilatih) 288 X 288, ukuran *input size* 576 X 576, dan nilai iterasi 10000. Skenario yang digunakan eksperimen sebagai berikut:

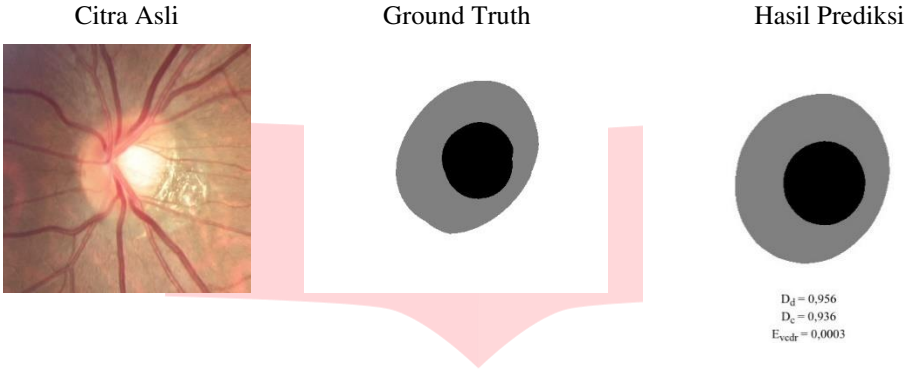
1. Skenario 1: Menggunakan dataset REFUGE dan *decoder* MLA. Dengan komposisi 800 citra latih dan 400 citra uji total 1200 citra yang digunakan.
2. Skenario 2: Menggunakan dataset REFUGE dan DRISHTI-GS1 dan *decoder* MLA saja. Dengan komposisi 800 citra latih dan 51 citra uji total 851 citra yang digunakan.
3. Skenario 3: Menggunakan *cross training* REFUGE dan DRISHTI-GS1 dengan DRISHTI-GS1 sebagai data uji dengan variasi bentuk *decoder* SETR. Dengan komposisi 850 citra latih dan 51 citra uji total 901 citra yang digunakan.

4. Skenario 4: Menggunakan *cross training* REFUGE dan DRISHTI-GS1 dengan REFUGE sebagai data uji dengan bentuk *decoder* yang mendapat hasil terbaik skenario 3. Dengan komposisi 850 citra latih dan 400 citra uji total 1150 citra yang digunakan.

Berikut ini hasil pengujian setiap skenario.

1. Skenario 1

skenario ini pengujian dilakukan dengan menggunakan REFUGE sebagai data latih, data validasi, dan data uji. Skenario ini menguji *decoder* bentuk MLA saja. saat pelatihan data selesai Cross nilai Cross Entropy Loss yang didapatkan 0.005 dan Dice Loss((Dice loss cup + Dice loss disc) / 2) 0.0013. Hasil pengukuran dapat dilihat tabel 3



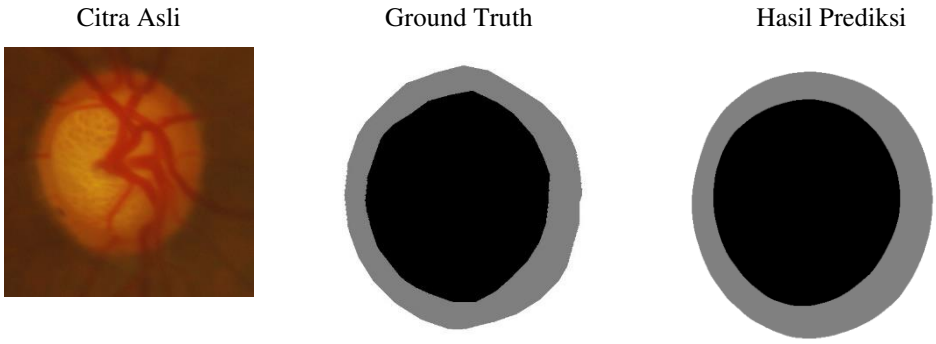
GAMBAR 13:
SALAH SATU PERBANDINGAN CITRA *FUNDUS* ASLI, ANOTASI *MASKS*(*GROUND TRUTH),* DAN HASIL PREDIKSI SETR DATA UJI REFUGE SKENARIO 1 DENGAN NILAI DICE SCORE DISC 0.956, DICE SCORE CUP 0.936, DAN VCDR ERROR 0.0003.

TABEL 3:
HASIL PENGUKURAN SKENARIO 1

Dataset	Dice Score Disc	Dice Score Cup	Average Dice	vCDR error
Validation (REFUGE)	0.955	0.878	0.917	0.016
Test (REFUGE)	0.960	0.892	0.926	0.017

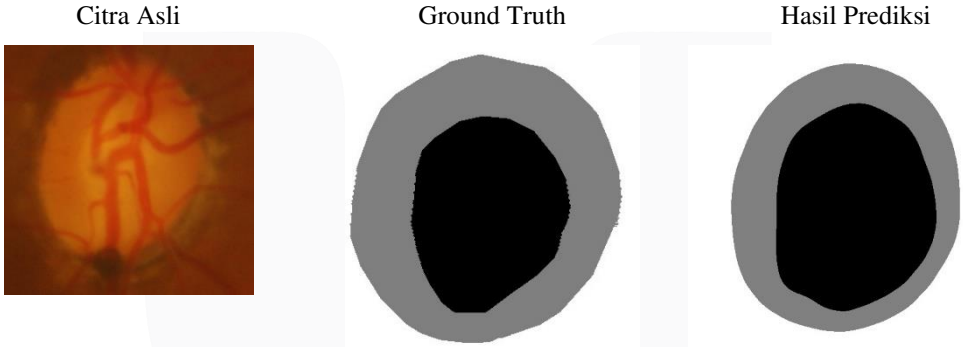
2. Skenario 2

skenario ini pengujian dilakukan dengan menggunakan model yang sebelumnya dilatih menggunakan training-set REFUGE dan validation-set REFUGE untuk dilakukan pengujian terhadap test-set DRISHTI-GS1. Skenario ini menguji decoder dengan bentuk MLA saja. saat pelatihan data selesai nilai Cross Entropy Loss yang didapatkan 0.007 dan Dice Loss((Dice loss cup + Dice loss disc) / 2) 0.0018. Untuk nilai *Dice score disc* mendapatkan nilai 0.670, *Dice score cup 0.672, Average dice* 0.671, dan *vCDR Error* 0.017.



GAMBAR 14:
SALAH SATU PERBANDINGAN ANTARA CITRA *FUNDUS* ASLI, ANOTASI *MASKS* (*GROUND TRUTH*), DAN HASIL PREDIKSI SETR DENGAN DATA LATIH REFUGE TERHADAP DATA UJI DRISHTI-GS1 SKENARIO 2 DENGAN DICE SCORE DISC 0.743, DICE SCORE CUP 0.810, DAN VCDR ERROR 0.059

ilustrasi **Gambar 14** yang merupakan salah satu citra dataset uji hasil prediksi menghasilkan D_d 0,670 , D_c 0,672 , dan E_{vcdr} 0,017. Selain contoh prediksi citra **Gambar 14**, terdapat pula contoh prediksi citra dengan nilai D_d dan D_c yang rendah.



GAMBAR 15:
SALAH SATU PERBANDINGAN ANTARA CITRA *FUNDUS* ASLI, ANOTASI *MASKS* (*GROUND TRUTH*), DAN HASIL PREDIKSI SETR DENGAN DATA LATIH REFUGE TERHADAP DATA UJI DRISHTI-GS1 SKENARIO 2 DENGAN *DICE SCORE* YANG RENDAH.

ilustrasi **Gambar 15** yang merupakan salah satu citra dataset uji hasil prediksi menghasilkan D_d 0,599 , D_c 0,476 , dan E_{vcdr} 0,547.

3. Skenario 3

skenario ini pengujian dilakukan dengan menggunakan model yang sudah dilatih menggunakan *training-set* dan *validation-set* REFUGE ditambah dengan *training-set* DRISHTI-GS1 untuk di uji *test-set* DRISHTI-GS1. Untuk uji validasinya menggunakan *training-set* REFUGE dan training-set DRSHITI-GS1 sebagai data latih untuk model dan menggunakan *validation-set* REFUGE sebagai validasinya. Selain itu skenario ini turut di lakukan uji coba terhadap varian *decoder* SETR yaitu menggunakan MLA, PUP, dan Naive. Nilai *Cross Entropy Loss* yang didapatkan dan *Dice Loss*((*Dice loss cup* + *Dice loss disc*) / 2) dilihat **Tabel 4.**

TABEL 4:
CROSS ENTROPY DAN DICE LOSS SIMULASI 3

Decoder	Cross Entropy Loss	Dice Loss

MLA	0.007	0.017
PUP	0.014	0.020
NAÏVE	0.010	0.015

Untuk hasil D_d, D_c, *Average dice* dan E_{vcdr} simulasi 3 dapat dilihat **Tabel 5** dan **Tabel 6.**

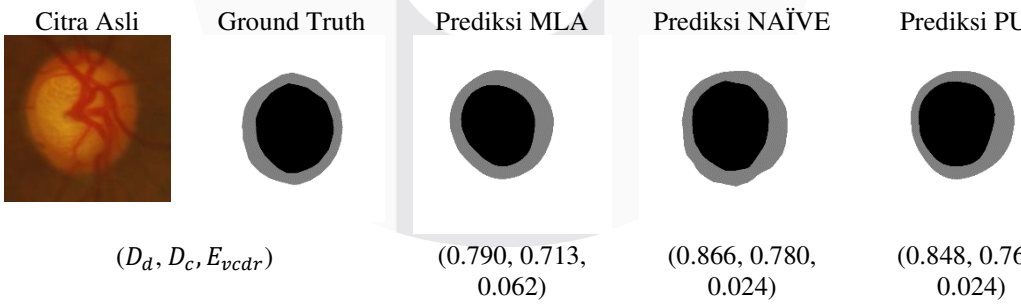
TABEL 5:
HASIL PENGUKURAN SKENARIO 3 DATA VALIDASI

Decoder	Dice Score Disc	Dice Score Cup	Average Dice	vCDR Error
MLA	0.951	0.877	0.914	0.018
PUP	0.944	0.870	0.907	0.019
NAÏVE	0.955	0.874	0.915	0.018

TABEL 6:
HASIL PENGUKURAN SKENARIO 3 DATA UJI

Decoder	Dice Score Disc	Dice Score Cup	Average Dice	vCDR Error
MLA	0.862	0.765	0.814	0.026
PUP	0.848	0.766	0.807	0.027
NAÏVE	0.866	0.780	0.823	0.024

Ilustrasi berikut ini akan memaparkan salah satu hasil prediksi *masks ground truth* dengan menggunakan 3 *decoder* yang dipaparkan

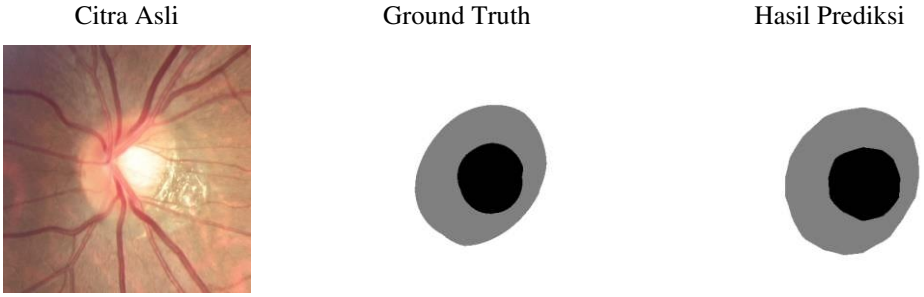


GAMBAR 16:
PERBANDINGAN CITRA *FUNDUS* ASLI, ANOTASI *MASKS* (*GROUND TRUTH*), DAN HASIL PREDIKSI SETR DECODER MLA, NAÏVE, DAN
PUP SIMULASI 3 SALAH SATU CITRA DATA UJI.

4. Skenario 4

skenario ini pengujian dilakukan dengan menggunakan model yang sudah dilatih menggunakan training-set dan validation-set REFUGE ditambah dengan training-set DRISHTI-GS1 untuk di uji test-set REFUGE. Untuk decoder yang digunakan decoder yang mempunyai Average Dice terbaik scenario 3 yaitu decoder NAÏVE. saat pelatihan data selesai nilai Cross Entropy Loss yang didapatkan 0.010 dan Dice Loss((Dice loss cup + Dice loss disc) / 2) 0.019. Untuk nilai *Dice*

score disc mendapatkan nilai 0.960, *Dice score cup* 0.890, *Average dice* 0.925, dan *vCDR Error* 0.017.



GAMBAR 17:
PERBANDINGAN CITRA *FUNDUS* ASLI, ANOTASI *MASKS* (*GROUND TRUTH*), DAN HASIL PREDIKSI SETR DECODER NAÏVE

ilustrasi **Gambar 17** yang merupakan salah satu citra dataset uji hasil prediksi menghasilkan D_d 0,959 , D_c 0,958 , dan E_{vcdr} 0,0003.

B. Analisis Hasil Pengujian

1. Dice Score berdasarkan bentuk decoder
   Berdasarkan hasil pengujian **Tabel 6** terlihat bentuk *decoder* NAIVE memberikan hasil yang lebih baik pengujian di data validasi atau pun data uji dengan rata rata *Dice Score* 0,915 untuk data validasi dan 0,823 data uji. Dengan demikian NAÏVE lebih baik dari 2 *decoder* lainnya scenario 3. Namun jika melihat hasil yang diberikan skenario 4, NAÏVE meski dengan jumlah data latih yang lebih banyak (850 citra), MLA **Tabel 3** (800 citra) mampu mengungguli NAIVE yang mampu mendapatkan rata-rata *Dice Score* 0,925 untuk data uji dan 0,915 untuk data validasi dengan rata-rata *Dice Score* 0,926 untuk data uji dan 0,917 untuk data validasi.

2. Pengaruh Cross-Train dataset terhadap Cross Dataset Evaluation
   Berdasarkan hasil pengujian skenario 2 dan hasil **Tabel 7** dapat terlihat jelas pengaruh *Cross-Train dataset* atau pencampuran 2 dataset yang berbeda untuk menguji dataset lainnya mampu memberikan hasil yang lebih baik. Melihat skenario 2 yang menggunakan *decoder* MLA dan hanya menggunakan REFUGE(training-set,validation-set) dataset sebagai data latih memberikan rata-rata Dice Score 0.671.Sementara hasil **Tabel 7** yang salah satu pengujiannya juga menggunakan *decoder* MLA dan dengan menggunakan kombinasi training-set DRISHTI-GS1 dan training-set,validation-set REFUGE memberikan rata-rata *Dice Score* 0.814.

V. KESIMPULAN

Berdasarkan pengujian penelitian tugas akhir ini *Segementation Transformer*s SETR berhasil melakukan segementasi *disc* dan *cup* dengan akurasi yang cukup baik dengan Dice Score tertinggi *cross train* dataset 0.866 untuk bagian *disc* dan 0.780 untuk bagian *cup* dengan decoder NAIVE. Namun tak dapat di abaikan fakta penggunaan SETR menggunakan jumlah paramater latih yang besar yaitu NAIVE(305.67 juta), MLA (310.57 juta), dan PUP (318.31 juta). Fakta *cross train* dataset memberikan perubahan nilai *Dice Score* yang signifikan, menunjukkan penggunaan *cross train* untuk melatih model agar mendapatkan peforma yang lebih baik dapat menjadi pertimbangan yang sangat baik.

Saran untuk pengembangan selanjutnya penggunaan SETR dengan jumlah paramater yang lebih sedikit agar dapat mengurangi *computational cost*.

REFERENSI

[1] S. Zheng *et al.*, “Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers,” 2020, doi: 10.1109/cvpr46437.2021.00681.

[2] A. Sevastopolsky, “Optic disc and cup segmentation methods for glaucoma detection with modification of U-Net convolutional neural network,” *Pattern Recognit. Image Anal.*, vol. 27, no. 3, pp. 618–624, 2017, doi: 10.1134/S1054661817030269.

[3] S. Sreng, N. Maneerat, K. Hamamoto, and K. Y. Win, "Deep learning for optic disc segmentation and glaucoma diagnosis on retinal images," *Appl. Sci.*, vol. 10, no. 14, 2020, doi: 10.3390/app10144916.

[4] S. Li, X. Sui, X. Luo, X. Xu, Y. Liu, and R. Goh, "Medical Image Segmentation using Squeeze-and-Expansion Transformers," pp. 807–815, 2021, doi: 10.24963/ijcai.2021/112.

[5] J. I. Orlando *et al.*, "REFUGE Challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs," *Med. Image Anal.*, vol. 59, 2020, doi: 10.1016/j.media.2019.101570.

[6] J. Sivaswamy, S. R. Krishnadas, and A. Chakravarty, "Dataset for the Assessment of Glaucoma from the Optic Nerve Head Analysis," *JSM Biomed Imaging Data Pap 2(1) 1004*, vol. 2, pp. 1–7, 2015.

[7] A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Nips, pp. 5999–6009, 2017.

[8] H. Fu, J. Cheng, Y. Xu, D. W. K. Wong, J. Liu, and X. Cao, "Joint Optic Disc and Cup Segmentation Based on Multi-Label Deep Network and Polar Transformation," *IEEE Trans. Med. Imaging*, vol. 37, no. 7, pp. 1597–1605, 2018, doi: 10.1109/TMI.2018.2791488.

[9] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2020, vol. 12346 LNCS, no. 7, pp. 213–229, doi: 10.1007/978-3-030-58452-8_13.

[10] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," 2020, [Online]. Available: http://arxiv.org/abs/2010.11929.

[11] S. Bakas *et al.*, "Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge," 2018, [Online]. Available: http://arxiv.org/abs/1811.02629.

[12] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor, "ImageNet-21K Pretraining for the Masses," pp. 1–20, 2021, [Online]. Available: http://arxiv.org/abs/2104.10972.