Evaluasi Invariansi Augmentasi pada CLIP dan DINOv2
Evaluating Augmentation Invariance in CLIP and DINOv2

Nahumi Nugrahaningsih[1)]

[1)]Jurusan Teknik Informatika, Universitas Palangka Raya
Jl. Hendrik Timang, Palangka Raya
[1)]nahumi@it.upr.ac.id

Abstrak

Vision foundation models (VFM) semakin banyak digunakan sebagai encoder visual dalam berbagai *task* computer vision. Meskipun demikian, stabilitas representasi visual yang dihasilkan oleh model pre-trained terhadap berbagai transformasi citra masih belum sepenuhnya dipahami. Penelitian ini menganalisis sensitivitas augmentasi pada dua VFM, yaitu CLIP ViT-B/32 dan DINOv2 ViT-B/14, ketika digunakan dalam kondisi frozen. Eksperimen dilakukan pada CIFAR-10 dengan lima jenis augmentasi citra: horizontal flip, random crop, color jitter, Gaussian blur, dan kombinasi augmentasi. Stabilitas representasi diukur menggunakan cosine similarity antara embedding citra asli dan citra hasil augmentasi serta intra-class embedding variance. Perbedaan antar model dianalisis menggunakan Wilcoxon signed-rank test dengan koreksi Benjamini–Hochberg false discovery rate, dan pengaruh jenis augmentasi diuji menggunakan Friedman test. Hasil menunjukkan bahwa CLIP secara konsisten memiliki augmentation invariance yang lebih tinggi dibandingkan DINOv2 pada seluruh kondisi augmentasi ($p < 0.001$). Perbedaan terbesar muncul pada Gaussian blur dengan effect size besar ($r = 0.866$), sedangkan perbedaan terkecil terjadi pada color jitter ($r = 0.139$). Hasil ini menunjukkan adanya *trade-off* antara kekayaan representasi dan stabilitas terhadap augmentasi pada vision foundation models dalam kondisi frozen. Temuan ini memberikan pemahaman empiris mengenai perilaku representasi visual pada dua model yang banyak digunakan dalam berbagai pipeline computer vision.

Kata kunci: Augmentasi citra, CLIP, DINOv2, Invariansi augmentasi, Model fondasi visi

Abstract

Vision foundation models (VFMs) are widely used as visual encoders in many computer vision tasks. However, the stability of visual representations produced by pre-trained models under image transformations remains insufficiently understood. This study analyzes augmentation sensitivity in two VFMs, CLIP ViT-B/32 and DINOv2 ViT-B/14, when used in a frozen setting. Experiments are conducted on CIFAR-10 using five augmentation types: horizontal flip, random crop, color jitter, Gaussian blur, and a combined augmentation condition. Representation stability is measured using cosine similarity between original and augmented embeddings and intra-class embedding variance. Statistical differences between models are evaluated using the Wilcoxon signed-rank test with Benjamini–Hochberg false discovery rate correction, while the overall effect of augmentation type is assessed using the Friedman test. The results show that CLIP consistently exhibits higher augmentation invariance than DINOv2 across all augmentation conditions ($p < 0.001$). The largest difference occurs under Gaussian blur with a large effect size ($r = 0.866$), while the smallest difference appears under color jitter ($r = 0.139$). These findings reveal a trade-off between representational richness and augmentation robustness in frozen vision foundation models and provide empirical insight into how two widely used models respond to common image transformations.

Keywords: *Augmentation invariance, CLIP, DINOv2, Image augmentation, Vision foundation*

DOI: https://doi.org/10.47111/jointecoms.v5i3 *Received: 01-09-2025*
Accepted: 10-09-2025

1. PENDAHULUAN

Augmentasi citra merupakan teknik yang banyak digunakan dalam *computer vision* untuk meningkatkan kemampuan generalisasi model. Berbagai transformasi seperti perubahan warna, pemotongan citra, atau pembalikan posisi sering diterapkan selama pelatihan agar model dapat mengenali objek dalam kondisi visual yang beragam. Teknik ini memungkinkan model belajar dari variasi data yang lebih luas tanpa perlu menambah jumlah citra secara eksplisit [1].

Perkembangan deep learning dalam beberapa tahun terakhir menghasilkan model dengan kemampuan representasi visual yang semakin kuat. Salah satu perkembangan penting adalah munculnya *Vision Foundation Models* (VFM) yang dilatih pada dataset berskala sangat besar dan dapat digunakan pada berbagai task pengenalan citra [2]. Model-model ini sering dimanfaatkan sebagai encoder yang tidak dilatih ulang pada *task downstream* karena mampu menghasilkan representasi visual yang bersifat umum.

Meskipun demikian, bagaimana model-model tersebut merespons transformasi citra tertentu masih menjadi pertanyaan yang penting. Salah satu cara untuk mempelajari perilaku tersebut adalah dengan menganalisis *augmentation invariance*, yaitu sejauh mana representasi visual tetap stabil ketika citra mengalami perubahan. Analisis ini menjadi relevan terutama ketika encoder digunakan dalam keadaan *frozen*, yaitu ketika bobot model hasil pre-training dipertahankan tetap dan tidak diperbarui selama eksperimen.

Penelitian ini menganalisis pengaruh berbagai augmentasi citra terhadap representasi visual yang dihasilkan oleh dua vision foundation models yang banyak digunakan, yaitu CLIP dan DINOv2. CLIP mempelajari representasi visual melalui pembelajaran kontras antara gambar dan teks dalam skala besar [3], sedangkan DINOv2 dilatih menggunakan pendekatan self-supervised berbasis patch yang menekankan hubungan struktural antar bagian citra [4]. Studi ini memberikan analisis empiris mengenai bagaimana kedua model tersebut merespons berbagai transformasi citra yang umum digunakan dalam *pipeline computer vision.*

2. TINJAUAN PUSTAKA

2.1 Augmentation Invariance dalam Self-Supervised Learning

Augmentation invariance merupakan konsep penting dalam pembelajaran representasi visual berbasis self-supervised. Dalam pendekatan ini, model dilatih agar menghasilkan representasi yang konsisten meskipun citra mengalami berbagai transformasi selama proses pelatihan. Dengan demikian, model diharapkan dapat mempertahankan informasi semantik utama meskipun terjadi perubahan visual pada citra masukan.

Sejumlah penelitian menunjukkan bahwa jenis augmentasi yang digunakan selama pelatihan dapat memengaruhi struktur representasi yang dipelajari model. Ericsson et al. [5] menunjukkan bahwa invariansi yang terbentuk pada tahap *pre-training* berpengaruh pada kemampuan representasi untuk ditransfer ke berbagai *task downstream*. Dengan demikian, strategi augmentasi tidak hanya meningkatkan performa pelatihan, tetapi juga ikut membentuk karakter representasi visual yang dihasilkan model.

Beberapa penelitian lain mencoba mempelajari invariansi dengan pendekatan yang lebih fleksibel. Chavhan et al. [6] mengusulkan metode *amortised invariance learning* yang menyesuaikan objektif invariansi selama proses pelatihan. Pendekatan lain menggabungkan pembelajaran *invariant* dan *equivariant* untuk mempertahankan informasi transformasi sekaligus menjaga stabilitas representasi visual [7].

2.2 Robustness Augmentasi pada Vision-Language Models

Selain pada *self-supervised learning*, stabilitas representasi terhadap augmentasi juga menjadi perhatian pada model *vision–language*. Model seperti CLIP menghasilkan representasi visual yang tidak hanya dipengaruhi oleh struktur citra, tetapi juga oleh hubungan semantik antara gambar dan teks yang dipelajari selama *pre-training*.

Beberapa penelitian mencoba meningkatkan stabilitas representasi pada model tersebut melalui berbagai strategi pelatihan. Cai et al. [8] memperkenalkan pendekatan yang menggunakan *augmented text prompts* untuk memisahkan informasi konten dari variasi gaya visual. Pendekatan

lain berfokus pada konsistensi representasi antar augmentasi dengan mendorong kesamaan *cosine similarity* antara berbagai tampilan citra yang diaugmentasi [9].

Penelitian lain menggabungkan pembelajaran *self-supervised* dan *weakly supervised* untuk menghasilkan representasi visual yang lebih stabil dan lebih mudah ditransfer ke berbagai *task* pengenalan citra [10].

2.3 Celah Penelitian

Meskipun berbagai penelitian telah membahas invariansi terhadap augmentasi, sebagian besar studi masih berfokus pada tahap pelatihan atau *fine-tuning* model. Analisis mengenai bagaimana representasi visual dari model *pre-trained* merespons transformasi citra ketika encoder digunakan tanpa pelatihan ulang masih relatif terbatas.

Selain itu, perbandingan langsung antara beberapa *vision foundation models* dalam kondisi eksperimen yang sama juga jarang dilaporkan. Studi yang secara sistematis mengevaluasi respons model terhadap berbagai jenis augmentasi dengan pendekatan empiris yang konsisten masih sedikit ditemukan.

Penelitian ini bertujuan mengisi celah tersebut dengan menganalisis pengaruh berbagai augmentasi citra terhadap representasi visual yang dihasilkan oleh dua *vision foundation models* yang banyak digunakan, yaitu CLIP dan DINOv2.

Untuk menjawab pertanyaan tersebut, penelitian ini merancang kerangka eksperimen yang memungkinkan analisis sistematis terhadap stabilitas representasi visual pada berbagai kondisi augmentasi citra. Bagian berikut menjelaskan dataset yang digunakan, jenis augmentasi yang diterapkan, serta metode evaluasi yang digunakan untuk menganalisis perubahan representasi visual yang dihasilkan model.

3. METODE PENELITIAN

3.1 Model

Penelitian ini mengevaluasi dua vision foundation models, yaitu CLIP ViT-B/32 dan DINOv2 ViT-B/14, yang digunakan dalam kondisi *frozen*. CLIP dilatih melalui pembelajaran kontras pada sekitar 400 juta pasangan gambar–teks, sehingga menghasilkan representasi visual yang selaras dengan konsep semantik dalam bahasa alami [3]. Sebaliknya, DINOv2 dilatih menggunakan pendekatan self-supervised berbasis patch dengan mekanisme distilasi pengetahuan untuk menghasilkan representasi visual yang kaya secara struktural [4]. Kedua model memiliki perbedaan pada ukuran patch dan objektif pelatihan. CLIP menggunakan patch berukuran 32 × 32 piksel, sedangkan DINOv2 menggunakan patch 14 × 14 piksel. Dalam penelitian ini seluruh bobot encoder dipertahankan tetap selama eksperimen sehingga representasi yang dihasilkan sepenuhnya berasal dari model hasil pre-training.

3.2 Dataset

Eksperimen dilakukan menggunakan dataset CIFAR-10, yang terdiri dari 60.000 citra RGB dengan resolusi 32 × 32 piksel yang terbagi ke dalam 10 kelas objek [11]. Untuk menjaga konsistensi eksperimen, diambil sampel terstratifikasi sebanyak 1.000 citra dari bagian *test split*, yaitu 100 citra per kelas, menggunakan *random seed* tetap (*seed = 42*). Pendekatan ini memastikan distribusi kelas tetap seimbang sekaligus menjaga reprodusibilitas eksperimen. Sebelum proses ekstraksi embedding, seluruh citra diubah ukurannya menjadi 224 × 224 piksel agar sesuai dengan ukuran input standar pada kedua model. Selanjutnya dilakukan normalisasi menggunakan parameter ImageNet normalization dengan nilai mean [0.485, 0.456, 0.406] dan standar deviasi [0.229, 0.224, 0.225].

3.3 Kondisi Augmentasi

Penelitian ini mengevaluasi enam kondisi augmentasi citra yang dirangkum pada Tabel 1. Setiap kondisi augmentasi diterapkan secara independen terhadap citra input sebelum proses ekstraksi embedding.

Tabel 1. Kondisi augmentasi citra yang digunakan dalam eksperimen.

No	Augmentasi	Parameter	Tujuan
0	Tanpa augmentasi (baseline)	—	Referensi embedding tanpa perturbasi
1	Horizontal Flip	p = 1.0	Menguji invariansi terhadap refleksi spasial
2	Random Crop	area 80%	Menguji sensitivitas terhadap pergeseran posisi objek
3	Color Jitter	brightness = 0.4, contrast = 0.4, saturation = 0.4, hue = 0.1	Menguji invariansi terhadap perubahan fotometrik
4	Gaussian Blur	σ = 2.0, kernel 5 × 5	Menguji sensitivitas terhadap hilangnya frekuensi tinggi
5	Combined (Flip + Color Jitter)	aplikasi berurutan	Menguji respons terhadap kombinasi transformasi

3.4 Ekstraksi Embedding

Untuk setiap citra, embedding diekstraksi dari kedua model pada seluruh kondisi augmentasi. Pada CLIP, embedding diambil dari CLS token projection pada encoder visual sehingga menghasilkan vektor berdimensi 512. Pada DINOv2, embedding diambil dari CLS token pada blok transformer terakhir dengan dimensi 768. Seluruh embedding kemudian dinormalisasi menggunakan L2 normalization sebelum dilakukan perhitungan metrik evaluasi.

3.5 Metrik Evaluasi

Penelitian ini menggunakan dua metrik utama untuk mengevaluasi stabilitas representasi visual. ***Cosine similarity*** dihitung antara embedding citra asli dan embedding citra hasil augmentasi. Untuk setiap kondisi augmentasi diperoleh 1.000 nilai cosine similarity. Nilai rata-rata yang lebih tinggi menunjukkan bahwa model lebih invariant terhadap transformasi citra.

Selain itu dihitung ***intra-class embedding variance***, yaitu rata-rata variasi embedding dalam kelas yang sama. Nilai yang lebih rendah menunjukkan bahwa distribusi embedding dalam kelas lebih kompak.

3.6 Analisis Statistik

Perbedaan antara kedua model dianalisis menggunakan Wilcoxon signed-rank test pada distribusi cosine similarity yang berpasangan. Uji ini digunakan karena distribusi cosine similarity tidak selalu mengikuti distribusi normal. Ukuran efek dihitung menggunakan

$$r = | Z | / \sqrt{N} \quad (1)$$

dengan interpretasi:

- r < 0.1 : sangat kecil
- ≤ r < 0.3 : kecil
- ≤ r < 0.5 : sedang
- r ≥ 0.5 : besar

Untuk mengendalikan kesalahan akibat pengujian berganda digunakan Benjamini–Hochberg false discovery rate (FDR). Selain itu pengaruh keseluruhan jenis augmentasi terhadap *cosine similarity* dalam setiap model dianalisis menggunakan Friedman test, yaitu alternatif non-parametrik dari repeated-measures ANOVA.

4. PEMBAHASAN

4.1 Hasil Empiris

Tabel 2 menampilkan rata-rata *cosine similarity*, standar deviasi, serta hasil uji statistik untuk perbandingan antara CLIP dan DINOv2 pada setiap kondisi augmentasi. Tabel 3 menyajikan

intra-class embedding variance, sedangkan Tabel 4 merangkum hasil Friedman test. Visualisasi hasil ditunjukkan pada Gambar 1 dan Gambar 2.

CLIP menghasilkan cosine similarity yang lebih tinggi dibandingkan DINOv2 pada seluruh kondisi augmentasi. Semua perbedaan antar model signifikan secara statistik ($p < 0.001$ setelah koreksi FDR). Pada horizontal flip, kedua model menunjukkan invariansi tinggi (CLIP: 0.992 ± 0.006; DINOv2: 0.971 ± 0.015) dengan effect size besar ($r = 0.853$). Perbedaan menjadi lebih jelas pada random crop (selisih = 0.081; $r = 0.825$) dan combined augmentation (selisih = 0.078; $r = 0.759$). Perbedaan terbesar muncul pada Gaussian blur, di mana cosine similarity CLIP tetap tinggi (0.876 ± 0.047) sementara DINOv2 turun tajam menjadi 0.317 ± 0.128. Kondisi ini menghasilkan effect size terbesar ($r = 0.866$). Sebaliknya, color jitter menghasilkan perbedaan kecil antara kedua model (selisih = 0.010; $r = 0.139$), meskipun tetap signifikan secara statistik. Hasil Friedman test menunjukkan bahwa jenis augmentasi berpengaruh signifikan terhadap cosine similarity pada kedua model (CLIP $\chi^2 = 3099.08$; DINOv2 $\chi^2 = 3299.85$; $p < 0.001$). Nilai χ^2 yang lebih besar pada DINOv2 menunjukkan sensitivitas yang lebih tinggi terhadap perubahan augmentasi.

Tabel 2. Rata-rata cosine similarity (± standar deviasi) antara embedding citra asli dan citra hasil augmentasi, beserta hasil uji Wilcoxon signed-rank (FDR-corrected).

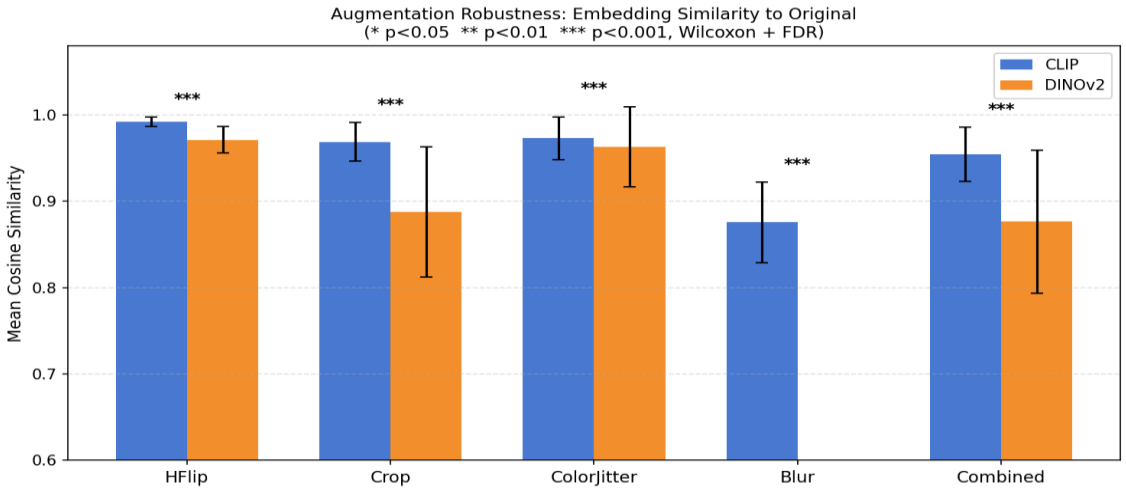
Augmentasi	CLIP (mean±std)	DINOv2 (mean±std)	Selisih	p (FDR)	Effect size r
Horizontal Flip	0.992 ± 0.006	0.971 ± 0.015	0.021	1.66×10^{-164}	0.853
Color Jitter	0.973 ± 0.024	0.963 ± 0.046	0.010	1.66×10^{-164}	0.139
Random Crop	0.969 ± 0.022	0.888 ± 0.075	0.081	1.66×10^{-164}	0.825
Combined	0.955 ± 0.031	0.876 ± 0.083	0.078	1.66×10^{-164}	0.759
Gaussian Blur	0.876 ± 0.047	0.317 ± 0.128	0.558	1.66×10^{-164}	0.866

Tabel 3. Rata-rata intra-class embedding variance ($\times10^{-4}$).

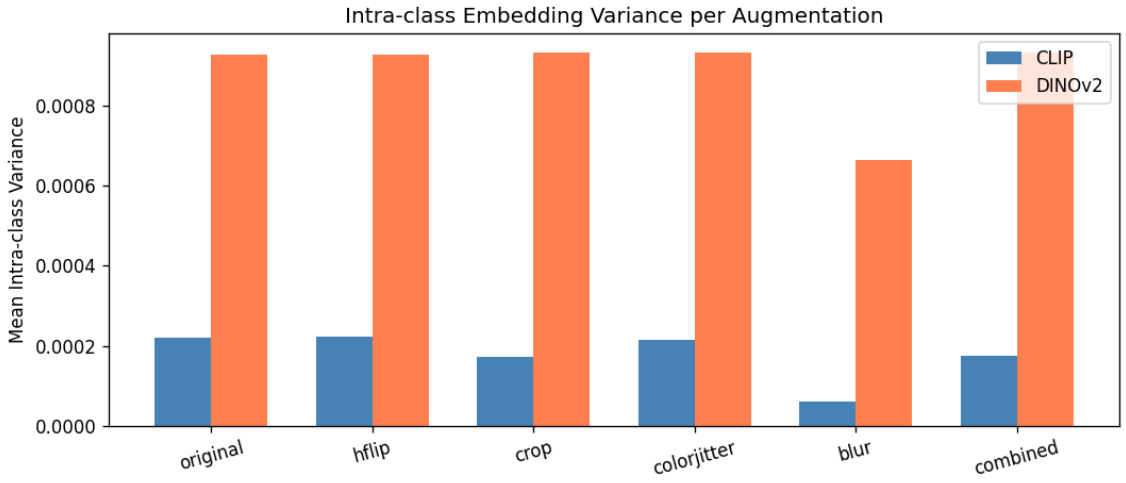
Kondisi	CLIP	DINOv2
Original	2.197	9.275
Horizontal Flip	2.216	9.283
Color Jitter	2.144	9.321
Random Crop	1.712	9.343
Combined	1.759	9.333
Gaussian Blur	0.607	6.648

Tabel 4. Hasil Friedman test untuk pengaruh jenis augmentasi terhadap cosine similarity.

Model	Friedman χ^2	p-value
CLIP	3099.08	< 0.001
DINOv2	3299.85	< 0.001



Gambar 1. Rata-rata cosine similarity antara embedding citra asli dan citra hasil augmentasi. Error bar menunjukkan ±1 standar deviasi. Penanda signifikansi menunjukkan hasil uji Wilcoxon setelah koreksi FDR (*** $p < 0.001$).



Gambar 2. Rata-rata intra-class embedding variance pada setiap kondisi augmentasi. DINOv2 menunjukkan variansi dasar yang lebih tinggi dibandingkan CLIP.

4.2 Pembahasan

Hasil eksperimen menunjukkan bahwa CLIP secara konsisten lebih stabil terhadap augmentasi dibandingkan DINOv2. Perbedaan ini dapat dijelaskan oleh perbedaan objektif pre-training kedua model.

CLIP dilatih menggunakan language-supervised contrastive learning, sehingga representasi visual cenderung menekankan konsep semantik tingkat tinggi. Pendekatan ini membuat embedding lebih tahan terhadap gangguan visual pada citra. Sebaliknya, DINOv2 menggunakan pendekatan self-supervised berbasis patch yang mempelajari hubungan struktural antar bagian citra. Representasi yang dihasilkan menyimpan informasi tekstur dan detail lokal yang lebih kaya, tetapi juga menjadi lebih sensitif terhadap perubahan struktur patch. Sensitivitas ini juga dilaporkan pada vision transformers berbasis patch dalam penelitian sebelumnya [12].

Perbedaan tersebut terlihat paling jelas pada Gaussian blur, yang menghilangkan komponen frekuensi tinggi pada citra. Karena DINOv2 bergantung pada informasi tekstur lokal, hilangnya

komponen ini menyebabkan penurunan stabilitas embedding yang besar. Hasil ini konsisten dengan penelitian yang menunjukkan bahwa vision transformers sensitif terhadap perubahan frekuensi citra [13], [14].

Sebaliknya, color jitter menghasilkan perbedaan kecil antara kedua model. Kedua model tetap mempertahankan cosine similarity yang tinggi. Hal ini kemungkinan karena DINOv2 dilatih menggunakan augmentasi warna selama pre-training, sehingga model telah belajar untuk mempertahankan invariansi terhadap perubahan fotometrik.

Secara keseluruhan, hasil ini menunjukkan adanya trade-off antara kekayaan representasi dan stabilitas terhadap augmentasi pada vision foundation models yang digunakan dalam kondisi frozen. CLIP menghasilkan representasi yang lebih stabil, sedangkan DINOv2 menghasilkan representasi yang lebih kaya secara struktural tetapi lebih sensitif terhadap transformasi citra.

5. KESIMPULAN

Penelitian ini menganalisis sensitivitas augmentasi pada dua vision foundation models dalam kondisi frozen, yaitu CLIP ViT-B/32 dan DINOv2 ViT-B/14, menggunakan metrik cosine similarity dan intra-class embedding variance pada lima kondisi augmentasi citra di dataset CIFAR-10. Signifikansi perbedaan dievaluasi menggunakan Wilcoxon signed-rank test dengan koreksi FDR dan Friedman test.

Hasil eksperimen menunjukkan bahwa CLIP secara konsisten memiliki augmentation invariance yang lebih tinggi dibandingkan DINOv2 pada seluruh kondisi augmentasi ($p < 0.001$). Perbedaan terbesar muncul pada Gaussian blur dengan effect size besar ($r = 0.866$), sedangkan perbedaan terkecil terjadi pada color jitter ($r = 0.139$).

Perbedaan ini berkaitan dengan objektif pre-training masing-masing model. CLIP menghasilkan representasi yang lebih stabil karena menekankan penyelarasan semantik gambar dan teks, sementara DINOv2 menghasilkan representasi yang lebih kaya secara struktural tetapi lebih sensitif terhadap perubahan tekstur dan frekuensi citra.

Secara keseluruhan, hasil ini menunjukkan adanya trade-off antara kekayaan representasi dan stabilitas terhadap augmentasi pada vision foundation models yang digunakan dalam kondisi frozen. CLIP lebih stabil untuk pipeline dengan variasi input yang tinggi, sedangkan DINOv2 lebih sesuai untuk aplikasi yang membutuhkan detail spasial yang lebih kaya.

DAFTAR PUSTAKA

[1] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *Journal of Big Data*, vol. 6, no. 60, 2019, doi: 10.1186/s40537-019-0197-0.
[2] R. Bommasani *et al.*, “On the Opportunities and Risks of Foundation Models,” 2021.
[3] A. Radford *et al.*, “Learning Transferable Visual Models From Natural Language Supervision,” in *Proceedings of the International Conference on Machine Learning*, 2021.
[4] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, and et al., “DINOv2: Learning Robust Visual Features without Supervision,” *Transactions on Machine Learning Research*, 2023.
[5] L. Ericsson, H. Gouk, and T. M. Hospedales, “Why Do Self-Supervised Models Transfer? On the Impact of Invariance on Downstream Tasks,” in *Proceedings of the British Machine Vision Conference*, 2022.
[6] R. Chavhan, J. Stuehmer, and C. Heggan, “Amortised Invariance Learning for Contrastive Self-Supervision,” in *Proceedings of the International Conference on Learning Representations*, 2023.
[7] X. Xu and J. Triesch, “CIPER: Combining Invariant and Equivariant Representations Using Contrastive and Predictive Learning,” in *Lecture Notes in Computer Science*, 2023.
[8] Y. Cai, Y. Liu, Z. Zhang, and J. Q. Shi, “CLAP: Isolating Content from Style Through Contrastive Learning with Augmented Prompts,” in *Lecture Notes in Computer Science*, 2025.
[9] M. Singha, A. Jha, and B. Banerjee, “GOPRO: Generate and Optimize Prompts in CLIP using Self-Supervised Learning,” in *Proceedings of the British Machine Vision Conference*, 2023.

[10] C. E. Bonhage, J. L. Mueller, A. D. Friederici, and C. J. Fiebach, "Combined eye tracking and fMRI reveals neural basis of linguistic predictions during sentence comprehension," *Cortex*, vol. 68, pp. 33–47, 2015.
[11] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," 2009.
[12] Y. Guo, D. Stutz, and B. Schiele, "Improving Robustness of Vision Transformers by Reducing Sensitivity to Patch Corruptions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
[13] M. Huang, W. Yu, and L. Zhang, "DF3Net: Dual Frequency Feature Fusion Network with Hierarchical Transformer for Image Inpainting," *Information Fusion*, 2024.
[14] M. F. Aslan, B. Aslan, and K. Sabanci, "Frequency-Domain Vision Transformers: Architectures, Applications, and Open Challenges," *Applied Sciences*, 2026.