

# **Deteksi Deepfake Video Berbasis VTT Vision Temporal Transformer**

---

Ikhlasul Amal

Magnolia Gina Ro'fataka Satriorini

Nilam Mufidah

Sisi Florensia

LOCALLY ROOTED ,  
GLOBALLY RESPECTED



ugm.ac.id

# Ancaman Deepfake yang Semakin Meningkat

Teknologi generatif (GANs) menciptakan video palsu dengan realisme yang mengkhawatirkan, mengancam integritas media dan kepercayaan publik.

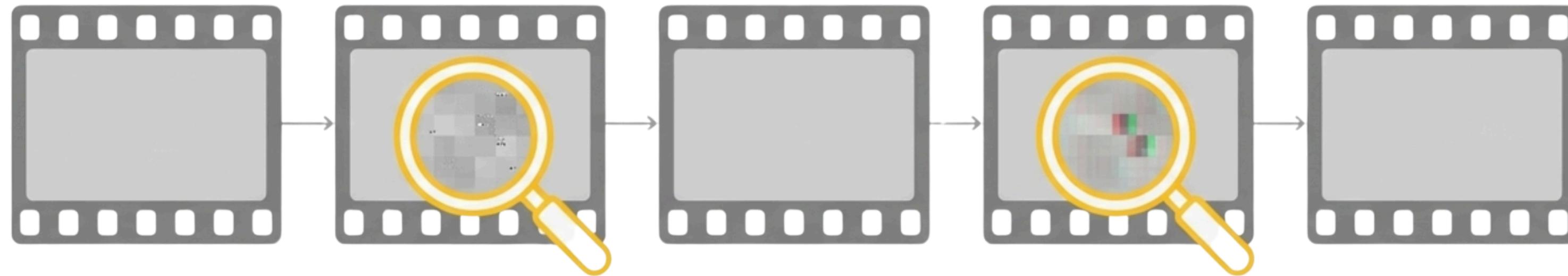
Kebutuhan akan mekanisme deteksi yang andal dan mampu mengidentifikasi artefak manipulasi yang halus menjadi sangat mendesak.



# Celah Deteksi Saat Ini: Fokus pada Artefak Spasial

Banyak metode deteksi kontemporer hanya berfokus pada analisis artefak spasial dalam satu frame video.

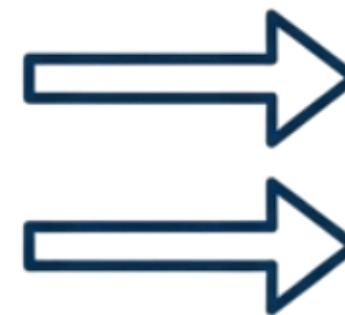
Padahal, kerentanan utama deepfake seringkali termanifestasi sebagai inkonsistensi temporal kegagalan mempertahankan konsistensi artefak fisik atau fisiologis antar-frame.



1. **Frame Awal:** Artefak tidak terlihat.
2. **Inkonsistensi Mikro:** Artefak halus (misalnya pikselasi) muncul dalam sekejap.
3. **Frame Berikutnya:** Artefak menghitang atau berubah bentuk.
4. **Pola Inkonsistensi:** Ketidakselarasan ini terjadi berulang kali di sepanjang video, menjadi petunjuk kunci.

# Pendekatan Solusi: Model LIPINC untuk Analisis Spatio-Temporal

Model **LIPINC** yang dirancang untuk secara simultan memodelkan hubungan spasial dan dependensi temporal



## 1. Vision Temporal Transformer (VTT)

Menganalisis fitur di dalam  
dan di antara frame.

## 2. Input Ganda (Residu Temporal)

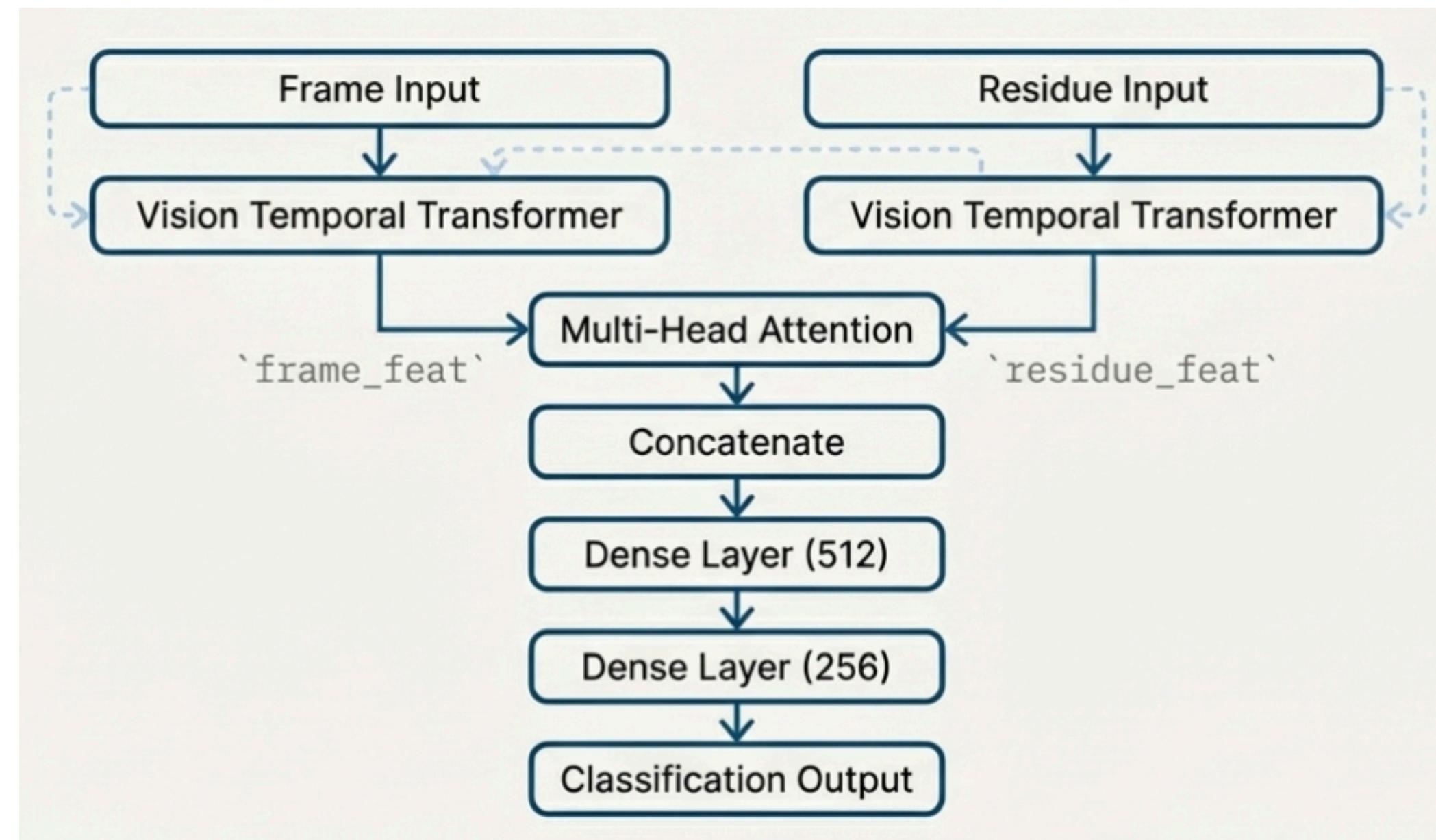
Memperkuat sinyal gerakan  
dan inkonsistensi  
transisional.

## 3. Fungsi Loss Ganda

Mengoptimalkan klasifikasi  
sekaligus mendorong  
konsistensi fitur.

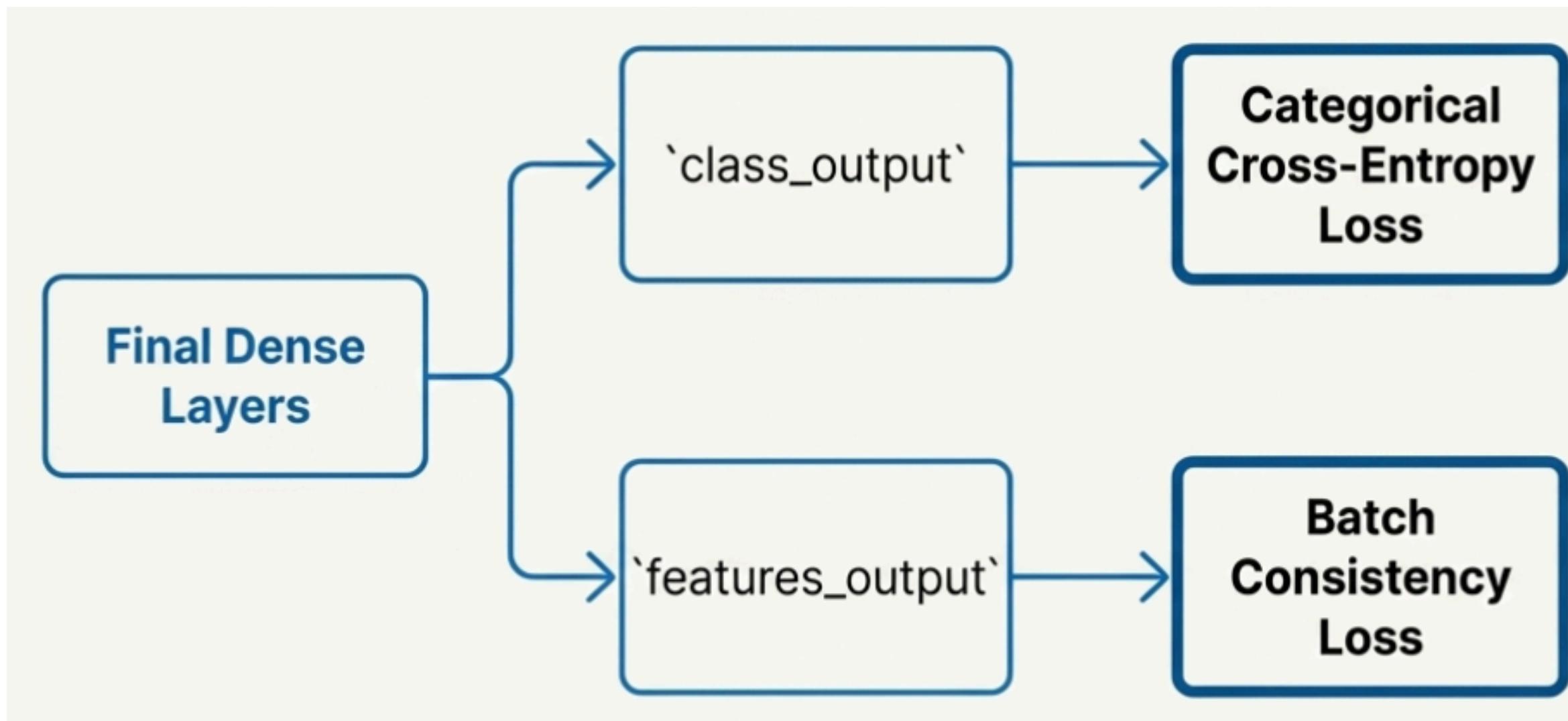
# Arsitektur Model LIPNIC

Kedua aliran input (frame dan residu) di proses secara paralel oleh blok Vision Temporal Transformer. Fitur yang dihasilkan kemudian digabungkan menggunakan *Multi-Head Attention* untuk menghasilkan klasifikasi akhir.



# Optimasi dengan Fungsi Loss Ganda

Model dioptimalkan menggunakan dua fungsi loss secara bersamaan untuk mencapai tujuan yang saling melengkapi



Berfokus pada tugas utama, mengklasifikasikan video sebagai **real** atau **fake** secara akurat.

Mendorong model untuk mempelajari representasi fitur yang konsisten dalam satu *batch*, sehingga lebih sensitif terhadap inkonsistensi yang melekat pada video *deepfake*.

# Desain Eksperimen

Kinerja model dievaluasi dalam **5 konfigurasi** berbeda untuk mengukur dampak dari skala data, keragaman dataset, dan teknik pra-pemrosesan.

**Dataset yang digunakan:** CelebDF V2, FaceForensics++ (FF++), Deepfake Detection (DFD), dan Deepfake Detection Challenge (DFDC).

**Metrik evaluasi:** ROC-AUC, Average Precision (AP), dan Intersection over Union (IoU).

Konfigurasi	Dataset Utama	Jumlah Dataset	Penggunaan MediaPipe
v1	CelebDF V2	1	Tidak
v2	FaceForensics++	1	Tidak
v3	CelebDF, FF++	2	Tidak
v4	FF++, DFD, DFDC	3	Tidak
v5	FaceForensics++	1	Ya



# Kombinasi Dataset

Dataset	V1	V2	V3	V4	V5
[R] CelebDF V2	590	0	590	0	0
[E] CelebDF V2	5,639*	0	5,639*	0	0
[R] FaceForensics++ (FF++)	0	1,000	1,000	1,000	1,000
[E] FaceForensics++ (FF++)	0	4,000*	4,000*	0	4,000*
[R] Deep Fake Detection (DFD)	0	0	0	363	0
[E] Deep Fake Detection (DFD)	0	0	0	3068*	0
[R] DFD Challenge (DFDC)	0	0	0	77	0
[E] DFD Challenge (DFDC)	0	0	0	0	0
TOTAL DATASET SIZE	1,180	2,000	3,180	2,880	2,000

[R] Real [F] Fake

\*Dilakukan *undersampling* agar membentuk dataset yang seimbang

# Hasil Eksperimen

Method	V1	V2	V3	V4	V5
<b>Data Counts (Training / Val / Test)</b>	Total: 1,180 Split: 826/177/177	Total: 2000 Split: 1600/300/300	Total: 3180 Split: 2226/477/477	Total: 2880 Split: 2016/432/432	Total: 2000 Split: 1600/300/300
<b>Recall Real Videos</b>	<b>0.89</b>	0.87	0.41	0.75	0.7
<b>Macro f1</b>	0.46	0.39	0.55	<b>0.85</b>	0.62
<b>AP</b>	0.5564	0.4656	0.5921	<b>0.8889</b>	0.6917
<b>ROC-AUC</b>	0.5582	0.4352	0.5921	<b>0.9256</b>	0.6866
<b>IOU</b>	0.1616	0.0859	0.4444	<b>0.7675</b>	0.4225
Link to Github Code	<a href="#">here</a>				

# Tes Inference

Inferensi model diuji pada himpunan data dunia nyata yang dikurasi dari artikel klarifikasi hoaks terkait deepfake di situs resmi **Komdigi** dan **Cekfakta Tempo**, di mana video palsu diekstraksi dari arsip berita tersebut.

Video asli diperoleh dari sumber rekaman sebelum dimanipulasi dan video autentik lain dengan publik figur yang sama yang telah diverifikasi bukan deepfake dari beragam media online.

Total terdiri dari 100 video, **50 video real** dan **50 video fake**.





# Hasil Tes Inference

Method	V1	V2	V3	V4	V5
Recall Real Videos	0.96	0.88	0.94	0.8	0.08
Macro f1	0.3243	0.3985	0.38	0.7494	0.4165
AP	0.5	0.4955	0.5	0.6944	0.5208
ROC-AUC	0.48	0.49	0.5	0.75	0.54
IOU	0	0.0893	0.0566	0.5833	0.5208
Latency	0.1855 s	0.2054 s	0.1819 s	0.1988 s	0.3072 s

# Kesimpulan

- 1 Superioritas Skala & Keragaman Data**  
Kinerja optimal (**ROC-AUC 0.9256**) hanya dicapai dengan melatih model pada tiga dataset berbeda (v4), membuktikan pentingnya generalisasi.
- 2 Efektivitas Arsitektur LIPINC**  
**Kombinasi VTT** (input residu, dan *consistency loss*) terbukti efektif dalam menangkap inkonsistensi *spatio-temporal* yang halus pada video deepfake.
- 3 Nilai Tambah Pra-pemrosesan**  
Penyesuaian wajah yang akurat (**MediaPipe**) dapat menjadi substitusi yang efektif untuk keragaman data ketika sumber data pelatihan terbatas.

*Terima Kasih*

