# CLaRE: CLIP with Latent Reconstruction Errors for Generated Face Detection

Udit Thakur, Asen Dotsinski, Aswin Krishna Mahadevan, Meher Changlani, Ioannis Kechagias, Hafeez Khan

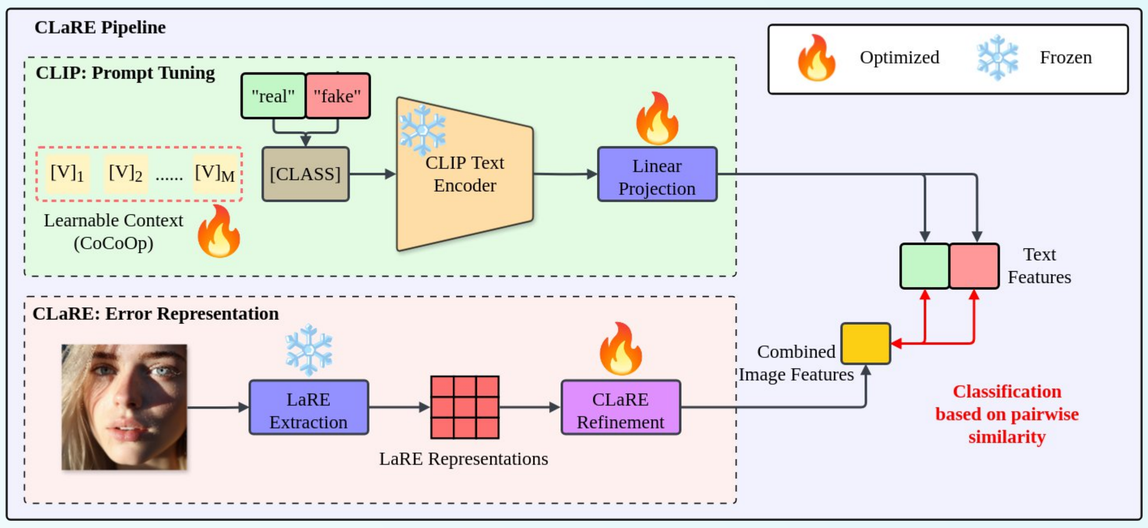
UNIVERSITY OF AMSTERDAM

**TL;DR: Detect generated faces with CLIP + reconstruction errors**
**CLaRE fuses prompt tuning in CLIP with LaRE, improving previous methods on generated face benchmarks!**

## Summary

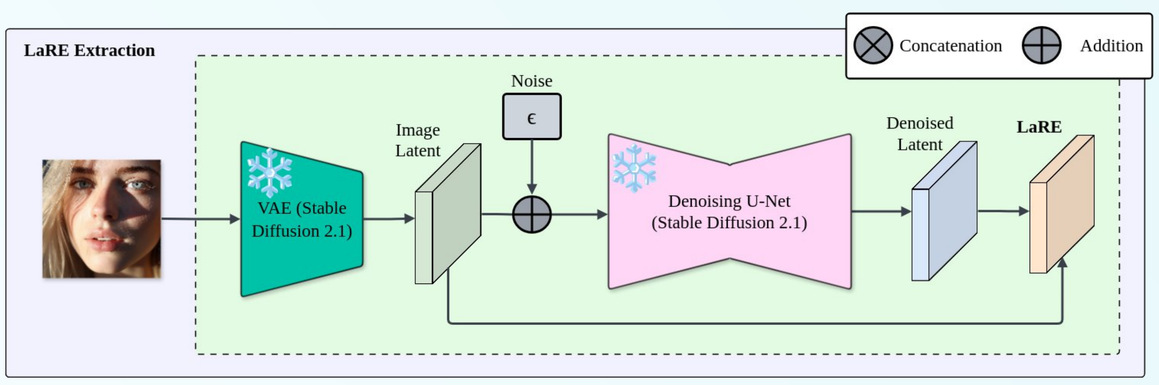
- CLIP with prompt tuning is a decent generated face detector, achieving high accuracy on almost all GAN-generated face images.
- However, it underperforms for diffusion-based generators, even with extensive training.
- CLaRE improves upon CLIP for facial diffusion images, by fusing Latent Reconstruction Errors (LaRE) into the model.

## 1. General Pipeline



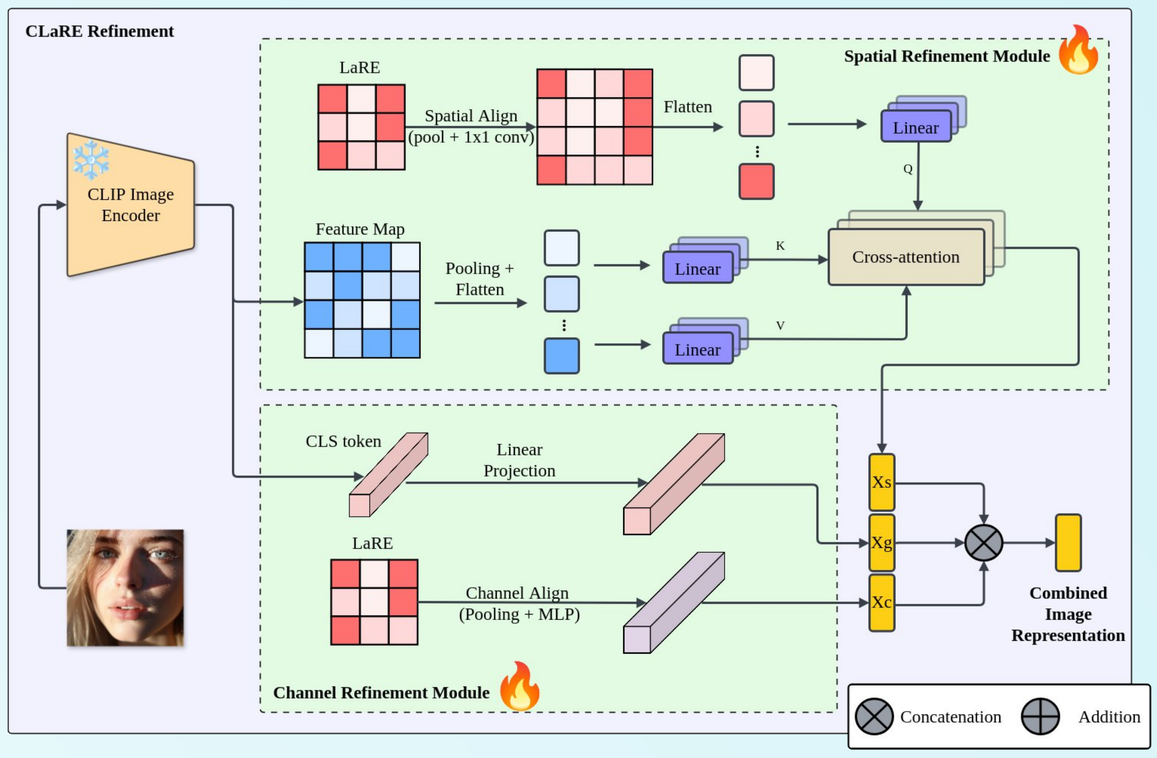

1. Textual features: We tune the text prompt using CoCoOp, encode it with the standard CLIP encoder and then project it linearly to match the size of the image features.
2. Image features: We first extract the Latent Reconstruction Error (LaRE) out of the image. Then, we process the image together with the LaRE representations to obtain learned image features that are more sensitive to diffusion artifacts.
3. Classification: Similarly to the original CLIP, we classify the image based on the representational similarity between the "real" and "fake" text features and the combined image features.

## 2. LaRE extraction




We use the same method as the original LaRE paper to extract the reconstruction errors, relying on Stable Diffusion 2.1. The difference between the denoised latent image and the original latent image is the latent reconstruction error (LaRE).

## 3. CLaRE Refinement




1. Spatial Refinement: The image is passed through the CLIP image encoder, producing patch features. These features are pooled and flattened. The LaRE representation is spatially aligned to image features through adaptive pooling and convolution, then flattened. We apply cross-attention between the image features and LaRE, producing the first part of our image representation.
2. CLS Channel Refinement: We use the CLS embedding of the image encoder as a global image representation, by linearly projecting it to the required size.
3. LaRE Channel Refinement: Similarly, we also pass the LaRE features separately, as they could be an important signal on their own. To align the dimensions, we apply pooling and a (learned) MLP.

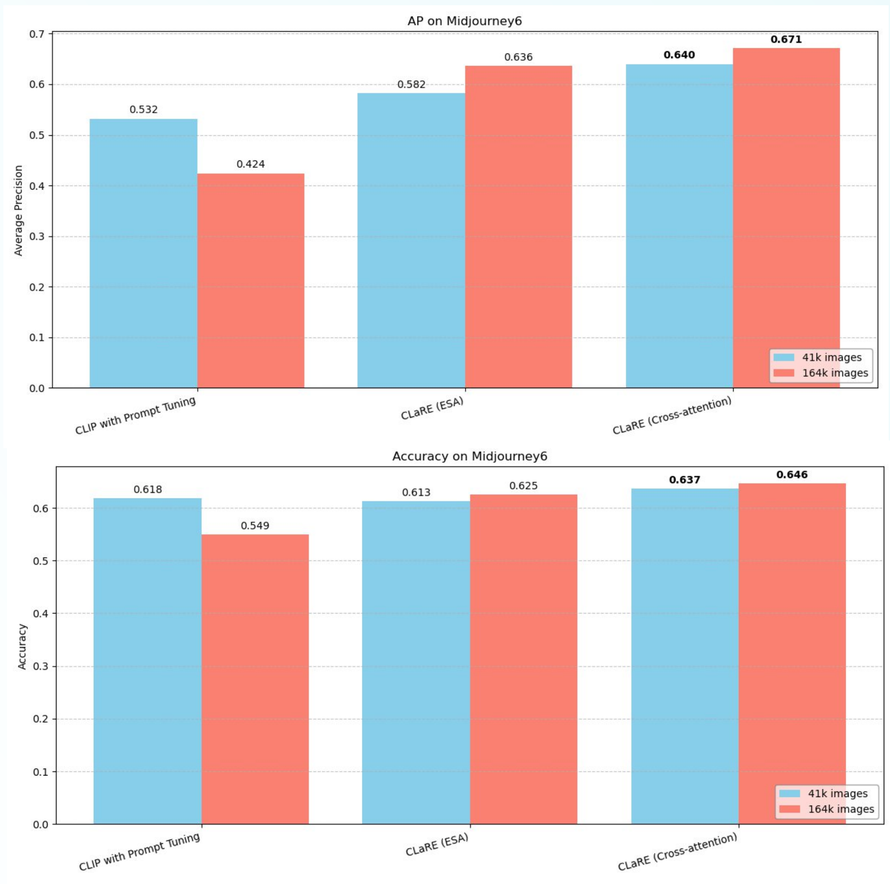
Finally, we concatenate the three representations from above to produce the combined image representation that is compared to the text features.

## 4. Dataset

We use a subset of DF-40 for training, particularly images generated by PixArt-α and SiT-XL/2 (diffusion-based), as well as VQGAN and StyleGAN-XL (GAN-based). The real training images come from FaceForensics++, while the real images for testing come from FF++ and CelebDF. We sample the datasets in a balanced manner for two types of training runs - one with 42k total images and one with 164k.

## 5. Results

- CLaRE outperforms CLIP with prompt tuning on Midjourney 6, in term of accuracy and especially AP, regardless of the size of the training dataset.
- CLIP with prompt tuning regresses in performance with more data, showing that it overfits to the training datasets and fails to generalize to Midjourney, unlike CLaRE.




- Both CLaRE and CLIP with prompt tuning saturate other EFS diffusion-based benchmarks, from which we tested Stable Diffusion 2.1 and RDDM.
- CLaRE outperforms CLIP with prompt tuning on CollabDiff, our other only diffusion-based benchmark, even though the task of face-editing is different from the one we trained for.
- CLaRE performs competitively on most other GAN-based benchmarks, although it lags behind significantly on WhichFaceIsReal and StarGAN V2. Given that the pipeline was optimized for diffusion-based images, the lower performance of CLaRE is to be expected.

| Model | Size | FE | | | | EFS | | | | FR |
|---|---|---|---|---|---|---|---|---|---|---|
| | | CollabDiff | StarGAN | StarGAN V2 | StyleCLIP | MidJourney6 | SD 2.1 | RDDM | WhichFaceIsReal | HeyGen |
| CLIP with Prompt Tuning | 41k | 0.653 | **0.985** | **0.864** | **0.959** | 0.532 | **0.995** | **0.990** | **0.605** | 0.953 |
| | 82k | 0.674 | 0.975 | 0.807 | 0.954 | 0.456 | - | - | 0.633 | 0.934 |
| | 164k | 0.701 | 0.968 | **0.854** | **0.954** | 0.424 | - | - | **0.698** | 0.954 |
| CLaRE (ESA) | 41k | **0.677** | 0.977 | 0.781 | 0.935 | 0.582 | - | - | 0.469 | 0.949 |
| | 164k | 0.659 | **0.976** | 0.792 | 0.935 | 0.636 | - | - | 0.500 | **0.958** |
| CLaRE (Cross-attention) | 41k | 0.616 | 0.953 | 0.782 | 0.901 | **0.640** | 0.990 | **0.990** | 0.509 | **0.959** |
| | 164k | **0.734** | 0.961 | 0.791 | 0.943 | **0.671** | **0.995** | 0.987 | 0.580 | 0.919 |

Table 1: Average Precision on various face-related datasets. CollabDiff, MidJourney6, SD 2.1 and RDDM are diffusion-based, the rest are GAN-based. **FE:** Face Editing, **EFS:** Entire Face Synthesis, **FR:** Face Reenactment

| Model | Size | FE | | | | EFS | | | | FR |
|---|---|---|---|---|---|---|---|---|---|---|
| | | CollabDiff | StarGAN | StarGAN V2 | StyleCLIP | MidJourney6 | SD 2.1 | RDDM | WhichFaceIsReal | HeyGen |
| CLIP with Prompt Tuning | 41k | 0.606 | **0.955** | **0.772** | **0.815** | 0.618 | **0.967** | **0.952** | **0.602** | **0.894** |
| | 82k | 0.614 | 0.929 | 0.718 | 0.812 | 0.588 | - | - | 0.615 | 0.830 |
| | 164k | 0.644 | 0.914 | **0.763** | **0.814** | 0.549 | - | - | **0.663** | **0.889** |
| CLaRE (ESA) | 41k | **0.626** | 0.926 | 0.694 | 0.781 | 0.613 | - | - | 0.531 | 0.884 |
| | 164k | 0.609 | **0.925** | 0.698 | 0.774 | 0.625 | - | - | 0.537 | 0.884 |
| CLaRE (Cross-attention) | 41k | 0.578 | 0.875 | 0.691 | 0.747 | **0.637** | 0.950 | **0.953** | 0.532 | 0.874 |
| | 164k | **0.674** | 0.891 | 0.696 | 0.804 | **0.646** | **0.968** | 0.946 | 0.583 | 0.853 |

Table 2: Accuracy comparison on various face-related datasets. CollabDiff, MidJourney6, SD 2.1 and RDDM are diffusion-based, the rest are GAN-based. **FE:** Face Editing, **EFS:** Entire Face Synthesis, **FR:** Face Reenactment