SAFIRE: Segment Any Forged Image Region

Myung-Joon Kwon[1*], Wonjun Lee[1*], Seung-Hun Nam[2], Minji Son[1], Changick Kim[1]

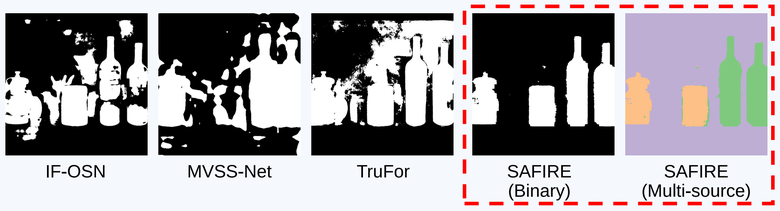
[1]Korea Advanced Institute of Science and Technology (KAIST) [2]NAVER WEBTOON AI, South Korea

Paper

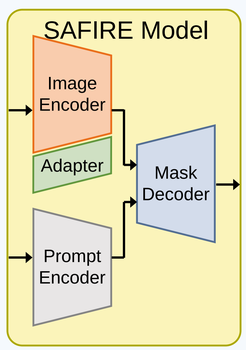
The 39th Annual AAAI Conference on Artificial Intelligence
FEBRUARY 25 – MARCH 4, 2025 PHILADELPHIA, PENNSYLVANIA, USA

Problem Formulation

- Image Forgery can create fake news, counterfeit evidence, and fraudulent microscopic images for paper mills.
- **Image Forgery Localization (IFL)**: Output the probability map of each pixel being forged.
- The forged image could consists of more than two sources. To trace the original image, we need multi-source partitioning.
- New Advanced IFL Task – **Multi-source Partitioning**: Partition the image into regions corresponding to their originating sources.



SAFIRE: Segment Any Forged Image Region



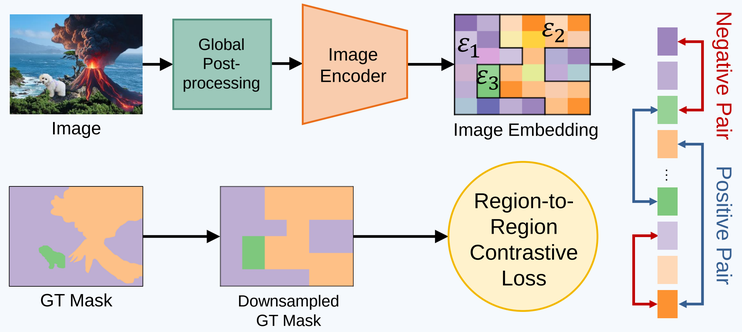
- Traditional IFL Methods: Train neural networks to label forged (1) and authentic (0) areas.
  - Cannot handle multi-source partitioning
  - Cannot deal with label agnosticity
- We view the IFL from a more fundamental perspective of *partitioning an image into distinct regions based on their origins*.
- We capitalize on Segment Anything Model (SAM)'s point prompting.
  - Each point prompt: Segments the source region containing itself.
- SAFIRE Model: SAM + Adapter to extract low-level feats.
- SAFIRE Framework: pretraining, training, and inference procedure.

Contributions

- Introduces a **new IFL task** for better interpretation of the image composition and easier further analysis.
- Proposes SAFIRE, a **novel IFL method** that uses **point prompting internally**. It is the first technique capable of multi-source partitioning.
- Achieves **SOTA performance** in both the traditional binary IFL and the new task.
- We construct and release a forgery **dataset** containing **images composed of multiple sources**.

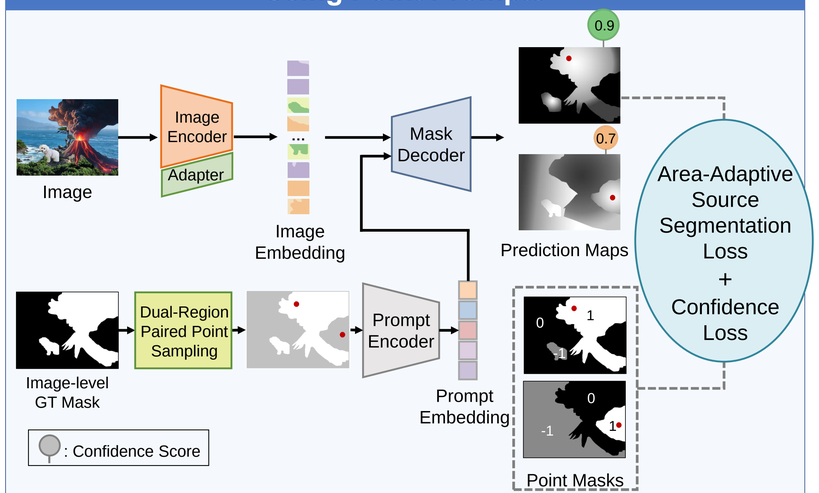
Code & Dataset
https://github.com/mjkwon2021/SAFIRE

Pretraining: Region-to-Region Contrastive Learning



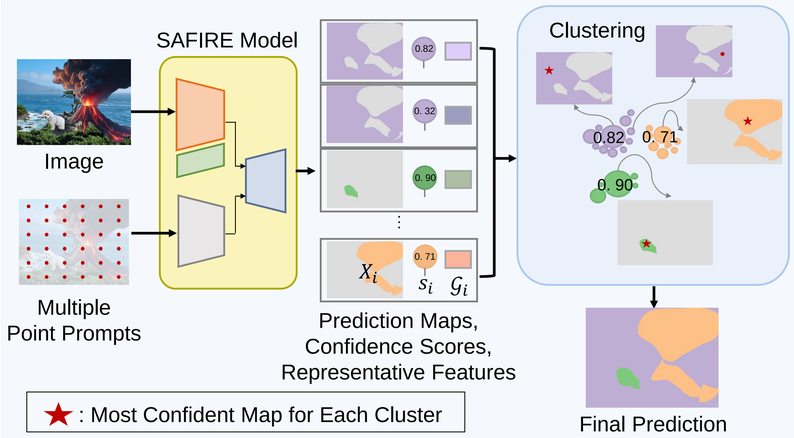
- Pretrain Image Encoder for effective source region partitioning.
- Aims to have embeddings from the same source region close together, while those from different source region are distanced.
- $\mathcal{L}_{R2R} = \frac{1}{|\mathcal{E}|}\sum_{i=1}^{r}\sum_{q\in\mathcal{E}_i} InfoNCE\left(q, \overline{\mathcal{E}_i\backslash\{q\}}, \mathcal{E}\backslash\mathcal{E}_i\right)$
- $InfoNCE(q,p,N) = -\log\left(\frac{exp\left(\frac{q\cdot p}{\tau}\right)}{exp\left(\frac{q\cdot p}{\tau}\right)+\sum_{n\in N} exp\left(\frac{q\cdot n}{\tau}\right)}\right)$

Training: Source Region Segmentation Using Point Prompts



- The goal of training is to make **each point prompt segment source region containing itself.**
  - Input: image + point prompt + GT mask
  - Output: prediction map + confidence score
- **Dual-Region Paired Point Sampling** is used to sample point prompts.
- Although SAFIRE can predict multi-source output, it can be trained with traditional forgery datasets – binary-labelled forged images.
  - To this end, we convert image-level mask to **Point Mask** depending on the point.
- **Area-Adaptive Source Segmentation Loss** guides the model to predict correct point mask.
- **Confidence Loss** is the pixel accuracy of predictions.

Inference: Multiple Point Aggregation



- Alongside the image, point prompts are provided as input to the model in a grid format internally.
- Each **Representative Features** are **clustered** and **highest confident masks** are used to yield the **final prediction**.

Visualizations