

ABSTRACT

This study was designed to address the challenges of webpage saliency prediction, particularly those arising from the structured and information-dense nature of webpage layouts and the limited availability of large-scale eye-tracking datasets. Unlike natural images, webpages organized visual content into hierarchical layouts composed of text blocks, images, navigation elements, and headers, where user attention was typically guided by specific goals such as reading, information seeking, or decision making. These characteristics reduced the effectiveness of saliency models trained primarily on natural-image datasets and highlighted the need for adaptation strategies that could generalize well under constrained data conditions.

To overcome these limitations, this research leveraged transfer learning to improve saliency prediction for webpage images by exploiting pretrained visual and contextual representations. Specifically, the decoder components of the TranSalNet architecture were fine-tuned using webpage saliency datasets, while the Transformer-based encoder was kept frozen to preserve its ability to capture global contextual dependencies. By retraining only the decoder layers, the model was encouraged to better interpret hierarchical webpage structures and to emphasize visually important elements such as text blocks, images, and headers, enabling effective domain adaptation with a relatively small amount of training data.

The proposed approach was evaluated on the FiWI and WebSaliency datasets using standard saliency metrics, including Normalized Scanpath Saliency (NSS), Correlation Coefficient (CC), Similarity (SIM), and Kullback–Leibler Divergence (KLDiv), with performance compared against the baseline TranSalNet model. On FiWI, decoder fine-tuning improved CC from 0.425 to 0.532, SIM from 0.359 to 0.426, and NSS from 1.291 to 1.639, while reducing KLDiv from 1.454 to 1.096. More substantial gains were observed on WebSaliency, where CC increased from 0.560 to 0.798, SIM from 0.574 to 0.707, NSS from 1.219 to 1.622, and KLDiv decreased markedly from 0.959 to 0.346.

The proposed method achieved particularly strong improvements on KLDiv, indicating improved alignment between predicted saliency distributions and ground-truth fixation distributions. A reduction of up to 63.9% on WebSaliency for the DenseNet-based TranSalNet showed that decoder-only fine-tuning effectively captured attention allocation at a distributional level. Overall, the DenseNet-based fine-tuned model achieved the best performance, demonstrating that decoder-only fine-tuning was an effective and computationally efficient strategy for adapting hybrid CNNs–Transformer saliency models to structured, text-heavy webpage domains.

Keywords : saliency, transformer, saliency prediction, webpage saliency dataset

INTISARI

Penelitian ini dirancang untuk menjawab tantangan dalam prediksi *saliency* pada citra halaman web, khususnya yang disebabkan oleh karakteristik tata letak halaman web yang terstruktur dan kaya informasi, serta keterbatasan ketersediaan data *eye-tracking* berskala besar. Berbeda dengan citra alami, halaman web menyusun konten visual dalam struktur hierarkis yang terdiri atas blok teks, gambar, elemen navigasi, dan header, di mana perhatian pengguna umumnya dipengaruhi oleh tujuan tertentu seperti membaca, mencari informasi, atau pengambilan keputusan. Karakteristik tersebut mengurangi efektivitas model *saliency* yang dilatih terutama pada dataset *natural images* dan menegaskan perlunya strategi adaptasi yang mampu melakukan generalisasi dengan baik dalam kondisi data yang terbatas.

Untuk mengatasi *gap* atau keterbatasan tersebut, penelitian ini menerapkan pendekatan *transfer learning* untuk meningkatkan performa prediksi *saliency* pada citra halaman web dengan memanfaatkan representasi visual dan kontekstual yang telah dipelajari sebelumnya. Secara khusus, *fine-tuning* dilakukan hanya pada komponen *decoder* arsitektur TranSalNet menggunakan dataset *saliency* halaman web, sementara encoder berbasis Transformer dipertahankan dalam kondisi dibekukan guna menjaga kemampuannya dalam menangkap dependensi kontekstual global. Dengan melatih ulang *decoder* saja, model menjadi lebih adaptif dalam menginterpretasikan struktur hierarkis halaman web dan menyoroti elemen visual penting, seperti blok teks, gambar, dan header, sehingga memungkinkan adaptasi domain yang efisien meskipun dengan data pelatihan yang terbatas.

Pendekatan yang diusulkan dievaluasi menggunakan dataset publik FiWI dan WebSaliency dengan metrik *saliency* standar, yaitu NSS, CC, SIM, dan KL-Div. Pada dataset FiWI, *fine-tuning decoder* meningkatkan CC dari 0,425 menjadi 0,532, SIM dari 0,359 menjadi 0,426, dan NSS dari 1,291 menjadi 1,639, sekaligus menurunkan KLDiv dari 1,454 menjadi 1,096. Peningkatan yang lebih signifikan diperoleh pada WebSaliency, dengan CC meningkat dari 0,560 menjadi 0,798, SIM dari 0,574 menjadi 0,707, NSS dari 1,219 menjadi 1,622, serta penurunan KLDiv yang besar dari 0,959 menjadi 0,346.

Hasil menunjukkan peningkatan yang sangat kuat pada metrik KLDiv, yang mencerminkan keselarasan yang lebih baik antara distribusi *saliency* prediksi dan distribusi fiksasi ground-truth. Penurunan KLDiv hingga 63,9% pada dataset WebSaliency untuk model TranSalNet berbasis DenseNet menunjukkan bahwa *fine-tuning decoder* efektif dalam menangkap alokasi perhatian pada tingkat distribusional. Secara keseluruhan, model hasil *fine-tuning* berbasis DenseNet mencapai kinerja terbaik, menegaskan bahwa pendekatan *fine-tuning decoder* saja merupakan strategi yang efisien secara komputasional untuk mengadaptasi model *saliency* hybrid CNN–Transformer pada domain webpage yang terstruktur dan kaya teks.

Kata kunci – saliency, transformer, saliency prediction, webpage saliency dataset