ABSTRAK

Sistem pengenalan wajah saat ini masih rentan terhadap *presentation attack*, yaitu upaya manipulasi identitas menggunakan media palsu seperti foto, video, atau masker. Salah satu pendekatan yang dikembangkan untuk mengatasi permasalahan tersebut adalah *face anti-spoofing multimodal* yang memanfaatkan citra RGB, *depth*, dan IR. Meskipun demikian, sebagian besar penelitian masih menggunakan arsitektur *backbone* yang sama untuk seluruh modalitas, padahal setiap modalitas memiliki karakteristik visual yang berbeda. Pendekatan tersebut berpotensi menghasilkan representasi fitur yang kurang optimal, sedangkan penentuan kombinasi *backbone* dan *cut block* secara manual memerlukan proses yang kompleks serta biaya komputasi yang tinggi. Oleh karena itu, penelitian ini bertujuan menemukan kombinasi *backbone* dan *cut block* yang paling sesuai untuk setiap modalitas secara otomatis guna meningkatkan kinerja deteksi.

Penelitian ini menggunakan *dataset benchmark* CASIA-SURF. Tahap akuisisi dan pra-pemrosesan data meliputi penyesuaian resolusi citra, augmentasi rotasi dan warna khusus pada modalitas RGB, serta konversi citra menjadi format *tensor*. Optimasi arsitektur dilakukan menggunakan *Bayesian Optimization* dengan algoritma *Tree-structured Parzen Estimator* (TPE). Selama proses pencarian, algoritma mengevaluasi kandidat arsitektur selama lima *epoch* untuk memilih *backbone* terbaik, yaitu *ResNet-18*, *ResNet-34*, atau *MobileNetV3-Large*, sekaligus menentukan *cut block* secara independen pada setiap modalitas berdasarkan nilai *validation loss*. Representasi fitur yang dihasilkan kemudian digabungkan menggunakan metode *Simple Additive Fusion*, sedangkan proses pelatihan model akhir dilakukan dengan *optimizer AdamW*.

Hasil penelitian menunjukkan bahwa algoritma TPE berhasil memperoleh konfigurasi arsitektur heterogen yang paling optimal, yaitu *MobileNetV3-Large* dengan *cut block* 9 pada modalitas RGB, *ResNet-34* dengan *cut block* 3 pada modalitas *depth*, serta *ResNet-18* dengan *cut block* 3 pada modalitas IR. Pada tahap pengujian, model hasil optimasi mencapai nilai *Average Classification Error Rate* (ACER) sebesar 0,2542% dan *Area Under the Curve* (AUC) sebesar 99,9923%. Kinerja tersebut lebih baik dibandingkan seluruh model *baseline* yang menggunakan arsitektur homogen. Hasil ini menunjukkan bahwa penggunaan *backbone* dan *cut block* yang disesuaikan dengan karakteristik masing-masing modalitas mampu meningkatkan performa sistem *face anti-spoofing* multimodal. Kontribusi utama penelitian ini adalah pengembangan kerangka optimasi arsitektur secara otomatis yang efektif untuk menentukan konfigurasi jaringan pada sistem *face anti-spoofing* multimodal.

Kata Kunci: *Face Anti-Spoofing*, *Multimodal*, *Bayesian Optimization*, Arsitektur *Backbone*, *Tree-structured Parzen Estimator*

ABSTRACT

Face recognition systems remain vulnerable to presentation attacks, in which identity is manipulated using spoofing media such as printed photographs, replayed videos, or masks. To address this issue, multimodal face anti-spoofing approaches have been developed by integrating RGB, depth, and infrared images. However, most existing studies still employ a homogeneous backbone architecture across all modalities, despite their distinct visual characteristics. Such an approach may produce suboptimal feature representations, while manually selecting the appropriate backbone architecture and cut block for each modality is computationally expensive and highly complex. Therefore, this study aims to automatically determine the most suitable backbone and cut block configuration for each modality to improve detection performance.

This study uses the public CASIA-SURF benchmark dataset. The data acquisition and preprocessing stages include image resizing, rotation and color augmentation applied specifically to the RGB modality, and conversion of raw images into tensor format. Architecture optimization is performed using Bayesian Optimization with the Tree-structured Parzen Estimator (TPE) algorithm. During the search process, candidate architectures are evaluated over five epochs to identify the most suitable backbone, namely ResNet-18, ResNet-34, or MobileNetV3-Large, while simultaneously determining the optimal cut block for each modality based on the validation loss. The resulting latent feature representations are then combined using the Simple Additive Fusion method, and the final model is trained using the AdamW optimizer.

The experimental results show that the TPE algorithm successfully identifies the optimal heterogeneous architecture, consisting of MobileNetV3-Large with cut block 9 for the RGB modality, ResNet-34 with cut block 3 for the depth modality, and ResNet-18 with cut block 3 for the infrared modality. The optimized model achieves an Average Classification Error Rate (ACER) of 0.2542% and an Area Under the Curve (AUC) of 99.9923% on the test set. This performance consistently outperforms all homogeneous backbone baseline models. These findings demonstrate that tailoring the backbone architecture and cut block to the characteristics of each modality leads to more effective multimodal face anti-spoofing. The main contribution of this study is the development of an automated architecture optimization framework for determining modality-specific network configurations in multimodal face anti-spoofing systems.

Keywords: *Face Anti-Spoofing*, *Multimodal*, *Bayesian Optimization*, *Backbone Architecture*, *Tree-structured Parzen Estimator*