

# Background-Robust Static Hand Gesture Recognition with Synthetic Scene Compositing: A Grouped Fivefold Ablation Across Five CNN Backbones

Cemal Tatlı

Ostim Technical University

Ramin Abbaszadi

`ramin.abbaszadi@ostimteknik.edu.tr`

Ostim Technical University

---

## Research Article

**Keywords:** Hand gesture recognition, Synthetic scene compositing, Grouped cross-validation, Transfer learning, Domain shift, Ablation study

**Posted Date:** October 26th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-10086694/v1>

**License:**   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

# Background-Robust Static Hand Gesture Recognition with Synthetic Scene Compositing: A Grouped Fivefold Ablation Across Five CNN Backbones

Cemal Tatlı<sup>1</sup> and Ramin Abbaszadi<sup>1\*</sup>

<sup>1</sup>Department of Software Engineering, Faculty of Engineering, OSTİM Technical University, Ankara, Türkiye.

\*Corresponding author(s). E-mail(s):

[ramin.abbaszadi@ostimteknik.edu.tr](mailto:ramin.abbaszadi@ostimteknik.edu.tr);

Contributing authors: [240801015@ostimteknik.edu.tr](mailto:240801015@ostimteknik.edu.tr);

## Abstract

Static hand gesture recognition models typically achieve high performance on small RGB datasets but exhibit reduced robustness as background complexity increases. This study investigates whether synthetic scene compositing can mitigate this vulnerability by employing a leakage-aware evaluation protocol. The proposed pipeline integrates alpha-matting-based foreground extraction with random scene compositing and is evaluated using a fivefold protocol that groups images at the image level, ensuring that augmented siblings from the same source image remain within the same fold. Five ImageNet-pretrained convolutional neural network (CNN) backbones are evaluated: EfficientNetV2B0, ResNet50, MobileNetV3Small, DenseNet121, and VGG16. Across 73 configurations and 363-fold runs from a single random seed, the complete pipeline increases mean validation accuracy from **48.0%** in the four-class no-synthesis baseline to **97.1%**. All 43 synthetic-baseline paired contrasts remain significant after Benjamini and Hochberg correction ( $q < 0.05$ , paired  $t$  test). The paired Wilcoxon test is constrained by the fivefold minimum attainable two-sided  $p$  value of  $\approx 0.0625$ . The protocol preserves transfer gap visibility: on a reserved HaGRID slice, direct transfer accuracy ranges from **66.2%** to **87.7%** across 67 synthetic augmented configurations, representing a mean decrease of **17.6** percentage points relative to internal validation (NO\_SYNTH excluded). An auxiliary MediaPipe Hands plus support vector machine (SVM) baseline achieves **59.2%** accuracy and **58.0%** macro  $F_1$ , primarily because landmark detection is

successful on only **37.3%** of that external slice. Progressive unfreezing underperforms standard fine-tuning in all paired comparisons (**18/18** pairs favor standard fine-tuning; **13/18** remain significant after BH FDR correction). Attention and spatial pooling heads do not outperform a plain global average pooling head. Code, raw per-fold JSON outputs, and reproduction scripts are available at <https://github.com/RageLog/bg-robust-rps>.

**Keywords:** Hand gesture recognition, Synthetic scene compositing, Grouped cross-validation, Transfer learning, Domain shift, Ablation study

## 1 Introduction

Vision-based static hand gesture recognition supports a broad spectrum of contactless interfaces, including gaming applications, assistive technologies, and touch-free controls in clinical environments. However, small RGB datasets frequently encourage reliance on spurious cues. Models may learn to recognize the room, lighting conditions, or camera configuration as readily as the intended hand shapes. Consequently, high internal evaluation scores may decline substantially when the visual context shifts.

The rock, paper, scissors task is revisited as it offers a concise framework for systematically examining data and model selection. Instead of focusing on a single prominent architecture, this study investigates three primary questions: the extent to which synthetic background compositing improves performance, whether these improvements are consistent across diverse ImageNet-pretrained convolutional neural network (CNN) backbones, and the degree to which these gains persist when models are transferred to an external benchmark without further tuning.

This paper makes five contributions:

1. We establish a fivefold protocol with image-level grouping, where the grouping rule is locally verifiable using the production filename schema ( $n = 39,150$  synthetic files). This approach prevents augmented siblings derived from the same parent image from being distributed across different folds.
2. We conduct a 73-configuration ablation study to isolate the effects of synthetic augmentation, backbone selection, head architecture, and fine-tuning strategy within a unified experimental framework.
3. We report aggregate scores alongside statistically corrected, fold-aligned comparisons across 363 runs, all initialized from a single random seed.
4. We explicitly quantify the residual performance gap between synthetic training and external transfer by directly evaluating on a reserved HaGRID subset.
5. We introduce an auxiliary baseline from an alternative modeling paradigm, specifically MediaPipe Hands combined with a support vector machine (SVM), to demonstrate the performance of a landmark-based classical pipeline on the same external four-class subset.

### *Paper organisation.*

The structure of the paper is as follows. Section 2 reviews the relevant literature and delineates the research gap. Section 3 details the dataset, synthetic pipeline, grouped evaluation protocol, and statistical procedures. Section 4 presents both internal and external results. Sections 5 through 7 discuss the main implications, limitations, and conclusions.

## 2 Related Work and Research Gap

### *Static gesture recognition with CNN backbones.*

Convolutional neural networks pretrained on large image corpora are the standard foundation for static gesture recognition. Recent studies demonstrate that compact architectures such as MobileNet achieve strong performance when datasets are visually constrained and backgrounds remain consistent [1]. However, many CNN studies continue to rely on predefined training and test splits, making it difficult to determine the extent to which reported performance is attributable to the model versus the data split. When augmentation and data splitting are intertwined, validation scores may be artificially inflated due to near-duplicate views from the same source image appearing in both training and test sets [2]. Another limitation is that many studies alter only one augmentation technique at a time, which obscures the source of observed performance improvements. There is a notable lack of leakage-aware comparisons across multiple backbones using a unified, grouped-image evaluation protocol.

### *Rock, paper, and scissors as a compact benchmark.*

The rock, paper, and scissors task serves as an effective small-vocabulary benchmark, as the three gestures possess a similar geometric structure yet remain easily distinguishable when the background is controlled. Recent studies that integrate MediaPipe hand landmarks with convolutional neural network (CNN) classification demonstrate strong internal performance; however, this approach does not address the more challenging issue of robustness to background variation [3]. Similarly, CNN-based studies using Kaggle-style rock, paper, scissors datasets frequently report near-perfect training or test accuracy but often lack rigorous external validation. Without evaluation on a reserved out-of-distribution dataset, it is unclear whether the model has learned the gesture itself or merely the visual patterns specific to the training data. Even in this constrained context, the literature lacks both a systematic augmentation ablation study and a comprehensive external transfer test using a larger, real-world benchmark.

### *Background augmentation and domain randomisation.*

Tobin et al. introduced domain randomization as a systematic approach for transferring policies from simulation to the real world by varying non-task-relevant visual attributes, including texture and lighting [4]. In the context of image classification, foreground compositing onto novel scenes represents a more targeted application of this concept. Awaluddin et al. combined alpha matting extraction with real background substitution for gesture recognition and demonstrated that the selection of the background pool significantly influences generalization [5]. More recently, Rahat et al.

investigated generative AI compositing as an additional source of training diversity and reported improvements over classical augmentation baselines at WACV 2025 [6]. However, the marginal contribution of individual compositing choices, such as augmentation ratio, edge softening, jitter, or background type, remains unclear when these factors are systematically evaluated across multiple backbones within a consistent experimental framework.

***Transfer learning and fine-tuning protocols.***

Residual networks demonstrated the importance of depth and skip connections for large-scale feature learning on ImageNet [7]. Subsequently, compound-scaled EfficientNet variants revealed that coordinated adjustments in resolution, depth, and width can enhance the balance between accuracy and parameter count [8]. In downstream fine-tuning, discriminative learning rate schedules—introduced in natural language processing via ULMFiT [9] and later popularized in computer vision through cyclical schedules [10]—are frequently assumed to benefit small datasets by preserving lower-level features. Progressive unfreezing operates under a similar rationale by gradually exposing additional backbone parameters. However, empirical evidence supporting these strategies for lightweight backbones and limited gesture vocabularies remains scarce. Negative results are rarely reported, resulting in limited understanding of when progressive fine-tuning is beneficial, neutral, or detrimental to performance.

***Statistical practice in computer vision ablation studies.***

Surveys on data augmentation have consistently argued that augmentation alters the effective training distribution and should be considered an integral component of experimental design rather than a minor detail [11]. Recent reviews of deep gesture recognition encompass a broad spectrum of convolutional neural network (CNN) architectures and application contexts [12]. However, many studies continue to rely solely on point estimates. In the absence of fold-aligned statistical analysis, a single training-and-test split may cause ordinary sampling variation to be misinterpreted as architectural superiority. Correction for multiple comparisons is even less common. Consequently, it is still rare for static gesture studies to report corrected significance across paired folds or to differentiate variation across folds from broader sources of uncertainty.

***Summary of the gap.***

In summary, the existing literature leaves three practical issues only partially addressed: preventing image-level augmentation leakage, isolating the contribution of individual compositing choices across multiple backbones, and reporting performance gains without obscuring the remaining gap to real-world data. This study is structured to address these three questions.

### 3 Methodology and Experimental Design

This section describes the dataset, the synthetic augmentation pipeline, the model training protocol, and the statistical framework used to interpret the ablations.

### 3.1 Dataset Construction

We combine four public rock, paper, and scissors collections [13–16] into a set of 7,830 unique parent images spanning three gesture classes. Table 1 summarises the class counts and the generated **none** class. Background scenes are drawn from the Indoor Scenes corpus [17]. A separate 600 image HaGRID slice (150 images per class) [18, 19] is reserved exclusively for external evaluation on a disjoint distribution (§3.9); it is never used for training, validation, or model selection.

**Table 1** Unique parent-image inventory (pre-synthesis) sourced from the combined RPS corpus. The **none** class is generated downstream from background scenes and is therefore not present in the raw gesture inventory.

| Class    | Image count | Share (%) |
|----------|-------------|-----------|
| paper    | 2,594       | 33.1      |
| rock     | 2,608       | 33.3      |
| scissors | 2,628       | 33.6      |
| Total    | 7,830       | 100.0     |

### 3.2 Synthetic Scene Augmentation

To reduce background dependence, we extract foreground hand regions from each parent image using a U<sup>2</sup>Net salient object segmenter [20] as implemented in **rembg**. Optional alpha matting softens the boundaries. The extracted foreground is then placed on a randomly selected background scene from §3.1. By default, the hand is centered with horizontal and vertical jitter of  $\pm 10\%$ , unless the **NO\_SHIFT** ablation disables this step. All composites are generated at the network input resolution of  $224 \times 224$ . Five composites are produced per parent image, unless the **RATIO\_1X** ablation reduces this to one. The pipeline follows the logic of domain randomisation [4] and is most directly comparable to the hybrid pipeline of [5] and the generative AI framework of [6]. As the merged dataset lacks pixel annotations, we assess segmentation quality using proxy measures: overlap with the MediaPipe hand box and edge contrast on the alpha mask. On a sample of 57 images, the U<sup>2</sup>Net mask achieves a mean proxy IoU of  $0.81 \pm 0.09$  and a mean edge contrast of 225.9, where higher values indicate sharper masks. We consider these values a sanity check rather than a calibrated segmentation score.

#### *Negative class construction.*

A dedicated **none** class represents the absence of a target gesture. Real negative samples are processed through the same compositing pipeline. To prevent the network from relying on background-only cues, every other **none** sample includes a hard distractor: either a real foreground image from the **hard negatives/** pool or a synthetic

skin-toned ellipse on the background. This approach is conceptually related to online hard example mining [21].

### 3.3 Grouped Image Cross Validation Protocol

We employ fivefold cross-validation with image-level grouping, utilizing scikit-learn’s stratified group splitter with shuffling enabled and a fixed random seed of 42. The group attribute corresponds to the parent raw image identifier, which is extracted from each synthetic filename by removing the class prefix and synthetic suffix. Consequently, all five augmented versions of a parent image are assigned to the same fold, ensuring that no augmented sibling of a training image appears in the corresponding validation set. We verified this property using the production filename schema for  $n = 39,150$  simulated synthetic files derived from 7,830 parent images, with all five folds exhibiting zero group overlap.

The Kaggle subsets used in this study do not provide subject-level metadata (e.g., user IDs). Thus, our grouping strategy prevents augmentation leakage but cannot guarantee the absence of subject leakage; this limitation is discussed in Section 6.

### 3.4 Backbone Architectures and Classification Heads

Five ImageNet-pretrained [22] convolutional neural network (CNN) backbones are evaluated: EfficientNetV2B0 [8], ResNet50 [7], MobileNetV3Small [23], DenseNet121 [24], and VGG16 [25]. Table 2 presents the trainable parameter footprint for the two backbones with available checkpoint metadata.

Consequently, trade-offs between model capacity and robustness are assessed using the complete five-backbone result grid, rather than relying solely on the partial checkpoint table.

The standard classification head consists of global average pooling, batch normalization, dropout ( $p = 0.3$ ), a fully connected layer with 256 units and  $L_2$  weight decay ( $10^{-4}$ ), a second batch normalization and dropout pair, and a four-class softmax layer. Two architectural ablations are also evaluated: a spatial pooling head that concatenates global average and max pooling, and a channel attention head based on squeeze and excitation.

**Table 2** Trainable-parameter footprint for the two retained backbone checkpoints with recoverable metadata (head excluded). Disk size varies by checkpoint format and is omitted for clarity.

| Backbone         | Trainable params (M) |
|------------------|----------------------|
| MobileNetV3Small | 1.09                 |
| EfficientNetV2B0 | 6.25                 |

### 3.5 Optimisation and Fine-Tuning Schedule

All models are initialized with ImageNet pretrained weights. Each fold uses a two-stage training schedule. In Phase 1 (warm up, 3 epochs), the backbone is frozen and only the head is trained. In Phase 2 (the remaining 17 epochs), the backbone is unfrozen and training continues at LR/10. All Batch Normalization layers remain in inference mode to preserve pretrained statistics.

We compare two unfreezing protocols. The standard protocol uses a single-stage full unfreeze. The progressive protocol unfreezes the top half of the backbone at LR/10 and the bottom half at LR/50, implementing a discriminative learning rate schedule. Training uses Adam [26] with clipnorm = 1.0, mixed precision [27] (float16 compute, float32 output), batch size 256, and  $224 \times 224$  inputs. We apply early stopping, learning rate reduction on plateau, best checkpoint selection based on validation loss, and NaN-termination callbacks. Data augmentation is limited to random horizontal flips and  $\pm 10\%$  brightness jitter. The random seed is set to 42 for the splitter, NumPy, TensorFlow, and data loader workers.

### 3.6 Factorial Ablation Design

A grid ablation is performed across four design factors: synthetic augmentation mode, backbone, classification head, and fine-tuning strategy. Each factor is varied independently, with the remaining factors held constant, to isolate and clarify the effect of each factor on performance.

The augmentation modes include **NO\_SYNTH** (no compositing), **REMBG\_ONLY** (transparent background), **RANDBG** (random noise background), **INDOOR** (indoor scene background), **GAN/STYLE.TRANSFER** (advanced generative backgrounds), and **FULL** (the proposed pipeline). Three sensitivity ablations are used to isolate compositing details: **RATIO\_1X** (factor 1 instead of 5), **NO\_SHIFT** (no jitter), and **NO\_ALPHA** (hard composite edges).

All five backbones are trained under each mode with the standard head and standard fine-tuning. This produces the main grid. Head and tuning ablations are evaluated on the representative **FULL/EfficientNetV2B0** cell.

### 3.7 Historical Comparator

We compare our grouped image protocol with an earlier single split version of the same pipeline, which achieved 99.91% on a single train/test partition. This earlier result is provided for historical reference only and is not included in the statistical procedure described in Section 3.11.

### 3.8 Auxiliary MediaPipe plus SVM Baseline

To establish a baseline from an alternative modeling tradition, a classical landmark pipeline is evaluated using MediaPipe Hands [28] in conjunction with a support vector machine (SVM) employing a radial basis function (RBF) kernel [29].

MediaPipe Tasks extracts a  $21 \times 3$  hand landmark vector. Images in which no hand is detected are encoded as a zero vector, consistent with the baseline implementation provided in the public code.

As the retained package does not provide a four-class `synthetic_full/train|val|test` split for this baseline, the SVM serves as an auxiliary diagnostic rather than as a paired member of the convolutional neural network (CNN) cross-validation grid. The SVM is trained on 7,830 raw rock-paper-scissors (RPS) gesture images and a deterministic seed 42 sample of 2,610 background images representing the `none` class. Evaluation is conducted on the retained external slice of 600 images (150 images per class).

### 3.9 Direct Transfer Evaluation on HaGRID

To assess genuine generalization, each trained model is evaluated directly on the reserved HaGRID slice. This ablation isolates the effect of model training by excluding fine-tuning, calibration, or threshold adjustment on HaGRID. External labels are mapped to the four-class taxonomy using a fixed dictionary. For instance, peace corresponds to scissors, palm to paper, fist to rock, and background only to none.

### 3.10 Temporal Smoothing for Deployment

During sequential webcam inference, the model output may be postprocessed using an  $N$ -frame majority filter and a per-class acceptance threshold  $\theta$ . A separate grid search is conducted over  $N \in \{1, 3, 5, 7, 10, 15\}$  and  $\theta \in [0.4, 0.9]$ , with each configuration evaluated using  $\text{score} = \text{accuracy} - \text{FPR}_{\text{none}}$ . The false-positive penalty is designed to discourage excessive activations in background frames. This analysis evaluates deployment stability and is presented separately from the primary cross-validation grid described in Section 3.6.

### 3.11 Statistical Analysis

For each pairwise comparison, the mean accuracy difference, paired Cohen’s  $d$ , paired  $t$  test  $p$  value, and paired Wilcoxon signed rank  $p$  value are reported, all computed across the same five grouped image folds.

Multiple comparisons within each family (synthetic versus no synthesis, backbone versus backbone, tuning strategy, head strategy) are controlled using the false discovery rate procedure of Benjamini and Hochberg at  $q < 0.05$  [30].

The minimum attainable two-sided  $p$  value of the paired Wilcoxon test at  $n=5$  folds is  $2/2^5 \approx 0.0625$ . Therefore, the paired  $t$  test is used as the primary inferential statistic, and the Shapiro and Wilk  $p$  value of the paired differences is reported as a normality check. Both tests are included in the supplementary CSV.

## 4 Results

The results are organized according to the sequence of the main research questions: internal grouped validation, followed by external transfer, and concluding with deployment-oriented diagnostics.

**Table 3** Main in-distribution results on the FULL synthetic-augmentation pipeline, standard classification head, image-grouped 5-fold cross-validation (mean $\pm$ std on validation accuracy, macro-F1 in %); cells with  $n < 5$  are flagged \*.

| Backbone         | Standard fine-tuning |          |         | Progressive fine-tuning |          |          |
|------------------|----------------------|----------|---------|-------------------------|----------|----------|
|                  | Acc. (%)             | Macro-F1 | Std $n$ | Acc. (%)                | Macro-F1 | Prog $n$ |
| EfficientNetV2B0 | 99.83 $\pm$ 0.05     | 99.83    | 5       | 93.21 $\pm$ 0.31        | 93.19    | 5        |
| MobileNetV3Small | 98.35 $\pm$ 0.05     | 98.35    | 5       | 82.86 $\pm$ 0.70        | 82.74    | 5        |
| ResNet50         | 99.93 $\pm$ 0.06     | 99.93    | 5       | 99.92 $\pm$ 0.01        | 99.92    | 5        |
| DenseNet121      | 99.96 $\pm$ 0.04     | 99.96    | 5       | 99.78 $\pm$ 0.02        | 99.78    | 5        |
| VGG16            | 99.93 $\pm$ 0.07     | 99.93    | 5       | 99.87 $\pm$ 0.03        | 99.87    | 5        |

## 4.1 Main Cross Validation Results

Table 3 presents validation accuracy and macro  $F_1$  scores for each backbone within the FULL pipeline. The results employ the standard classification head and both fine-tuning protocols, utilizing fivefold cross-validation and image-level grouping (mean $\pm$ SD). Figure 1 compares the FULL accuracy of each backbone to its corresponding NO\_SYNTH baseline.

With standard fine-tuning, four backbones achieve accuracy greater than 99.8%. The lightweight MobileNetV3Small also maintains accuracy above 98%. The primary internal performance difference is observed between models trained with scene synthesis and those trained without, rather than between different backbone families.

## 4.2 Effect of the Synthetic Pipeline

Aggregated results across all backbones, heads, and fine-tuning strategies indicate that the synthetic modes cluster between 96.6% and 97.9%, while the NO\_SYNTH ablation decreases to 48.0% (Table 4, Figure 2).

Table 5 summarizes the backbone-specific paired contrasts between each synthetic setting and NO\_SYNTH. All 43 comparisons remain significant after the Benjamini and Hochberg correction at  $q < 0.05$ , with mean improvements ranging from +43 to +57 percentage points and large positive Cohen’s  $d$  values.

## 4.3 Standard and Progressive Fine-Tuning

Table 6 presents the paired contrasts for each cell between standard and progressive fine-tuning. Among the 18 available pairs, standard fine-tuning is favored in all cases, with 13 pairs remaining significant following the Benjamini-Hochberg correction. The effect is most pronounced for EfficientNetV2B0 and MobileNetV3Small, where progressive unfreezing reduces mean accuracy by 5 to 15 percentage points. Figure 3 illustrates each paired difference.

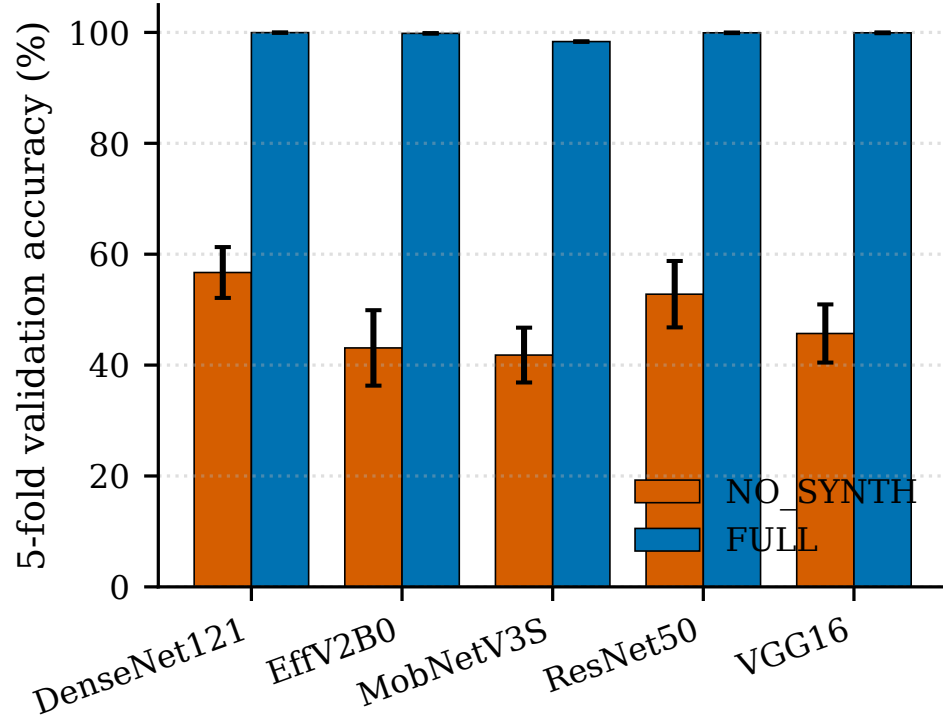


**Fig. 1** Validation accuracy of each backbone under the proposed FULL pipeline (blue) versus the NO\_SYNT baseline (orange) under standard fine-tuning; error bars show one standard deviation. The auxiliary MediaPipe Hands plus SVM baseline does not share this internal protocol; its external accuracy appears in Figure 7 for reference.

**Table 4** Aggregate impact of each augmentation/ablation mode, pooled across all backbones, heads, and tuning strategies (mean $\pm$ std accuracy in % across all available folds).

| Ablation mode  | Fold runs | Accuracy (%)     |
|----------------|-----------|------------------|
| NO_ALPHA       | 25        | 97.85 $\pm$ 5.11 |
| REMBG_ONLY     | 25        | 97.83 $\pm$ 4.20 |
| INDOOR         | 50        | 97.37 $\pm$ 5.36 |
| GAN            | 50        | 97.33 $\pm$ 5.28 |
| RANDBG         | 48        | 97.28 $\pm$ 5.46 |
| STYLE_TRANSFER | 25        | 97.27 $\pm$ 5.23 |
| RATIO_1X       | 25        | 97.23 $\pm$ 5.93 |
| FULL           | 65        | 97.14 $\pm$ 4.84 |
| NO_SHIFT       | 25        | 96.61 $\pm$ 6.06 |
| NO_SYNT        | 25        | 48.02 $\pm$ 7.78 |

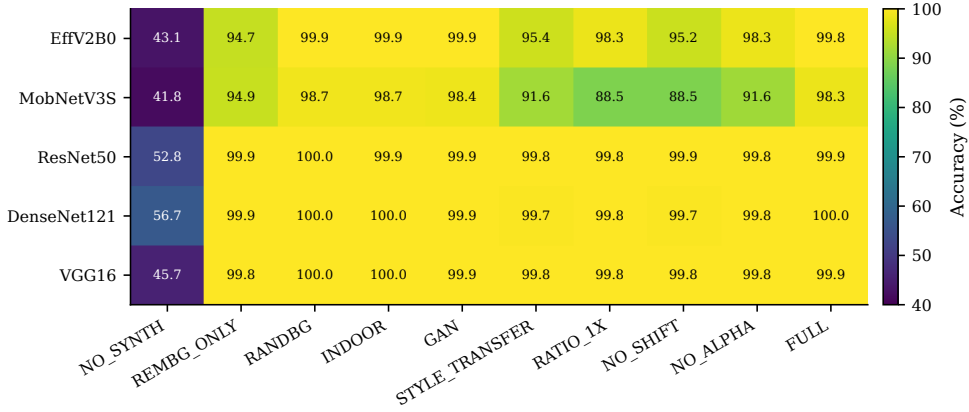


**Fig. 2** Mean validation accuracy by backbone × ablation mode (standard head, standard fine-tuning, fivefold mean). Only the **NO\_SYNTH** column collapses below 60%. The classical MediaPipe plus SVM baseline cannot be placed on the same axis because the augmentation modes do not apply to it; its external accuracy is reported in Table 11 and Figures 7 and 8.

**Table 5** Effect of the synthetic-augmentation pipeline against the no-augmentation baseline (**NO\_SYNTH**), aggregated across augmented modes per backbone. “Sig.” counts pairs with BH-FDR  $q < 0.05$  on the paired  $t$ -test. The **NO\_SYNTH** baseline collapses to  $\approx 45$ – $57\%$  accuracy in our 4-class setup, whereas every augmented mode reaches at least 80%. Wilcoxon results at the  $n=5$  two-sided floor ( $\approx 0.0625$ ) are reported in the supplementary CSV.

| Backbone         | $n$ pairs | Min $\Delta$ (pp) | Max $\Delta$ (pp) | Mean $\Delta$ (pp) | Mean $g$ | Sig. ( $q < .05$ ) |
|------------------|-----------|-------------------|-------------------|--------------------|----------|--------------------|
| DenseNet121      | 8         | +42.96            | +43.27            | +43.13             | +7.52    | 8/8                |
| EfficientNetV2B0 | 8         | +51.63            | +56.79            | +54.58             | +5.86    | 8/8                |
| MobileNetV3Small | 9         | +46.66            | +56.87            | +52.53             | +6.27    | 9/9                |
| ResNet50         | 9         | +47.01            | +47.19            | +47.09             | +6.29    | 9/9                |
| VGG16            | 9         | +54.11            | +54.27            | +54.18             | +8.25    | 9/9                |

#### 4.4 Effect of the Classification Head

Table 7 presents a comparison between the standard global average pooling head and two alternative approaches on the EfficientNetV2B0 **FULL** cell during standard fine-tuning. The standard head significantly outperforms both the attention and spatial pooling heads ( $q < 0.05$ ).

#### 4.5 Per Class Behaviour in Internal Validation

Prior to analyzing external transfer, the error structure of the best internal cell for each backbone is examined. Figure 4 presents the confusion matrix corresponding to each backbone. In each panel, the backbone’s optimal configuration among the complete fold cells is utilized. DenseNet121 and EfficientNetV2B0 achieve their highest performance under the **INDOOR** ablation, while the remaining three backbones peak under

**Table 6** Effect of progressive vs. standard fine-tuning, paired across the same 5 image-grouped folds (per-pair test: paired  $t$ -test;  $^\dagger$  denotes BH-FDR adjusted  $q < 0.05$  within this family). Positive  $\Delta$  means standard fine-tuning is better. The two-sided minimum achievable  $p$ -value for paired Wilcoxon at  $n=5$  is  $\approx 0.0625$ , so the  $t$ -test is used as the primary inferential statistic; full Wilcoxon results are in the supplementary CSV.

| Backbone         | Ablation | Standard (%) | Progressive (%) | $\Delta$ (pp) | Hedges' $g$ | $p$ ( $t$ )       |
|------------------|----------|--------------|-----------------|---------------|-------------|-------------------|
| DenseNet121      | FULL     | 99.96        | 99.78           | +0.18         | +7.66       | 0.0000 $^\dagger$ |
| DenseNet121      | INDOOR   | 99.96        | 99.80           | +0.16         | +4.38       | 0.0003 $^\dagger$ |
| DenseNet121      | GAN      | 99.86        | 99.76           | +0.10         | +1.23       | 0.0262 $^\dagger$ |
| EfficientNetV2B0 | FULL     | 99.83        | 93.21           | +6.62         | +15.85      | 0.0000 $^\dagger$ |
| EfficientNetV2B0 | INDOOR   | 99.88        | 93.05           | +6.83         | +15.42      | 0.0000 $^\dagger$ |
| EfficientNetV2B0 | GAN      | 99.85        | 92.90           | +6.95         | +20.11      | 0.0000 $^\dagger$ |
| MobileNetV3Small | GAN      | 98.36        | 82.93           | +15.44        | +31.99      | 0.0000 $^\dagger$ |
| MobileNetV3Small | RANDBG   | 98.68        | 82.63           | +16.05        | +16.38      | 0.0000 $^\dagger$ |
| MobileNetV3Small | FULL     | 98.35        | 82.86           | +15.48        | +17.91      | 0.0000 $^\dagger$ |
| MobileNetV3Small | INDOOR   | 98.66        | 82.68           | +15.99        | +22.26      | 0.0000 $^\dagger$ |
| ResNet50         | FULL     | 99.93        | 99.92           | +0.01         | +0.08       | 0.8303            |
| ResNet50         | INDOOR   | 99.93        | 99.91           | +0.02         | +0.30       | 0.4543            |
| ResNet50         | RANDBG   | 99.97        | 99.96           | +0.02         | +0.42       | 0.3046            |
| ResNet50         | GAN      | 99.93        | 99.91           | +0.02         | +0.54       | 0.2080            |
| VGG16            | RANDBG   | 99.96        | 99.89           | +0.08         | +2.42       | 0.0025 $^\dagger$ |
| VGG16            | FULL     | 99.93        | 99.87           | +0.06         | +0.60       | 0.1678            |
| VGG16            | INDOOR   | 99.96        | 99.86           | +0.09         | +1.94       | 0.0056 $^\dagger$ |
| VGG16            | GAN      | 99.91        | 99.85           | +0.06         | +1.24       | 0.0256 $^\dagger$ |

**Table 7** Classification-head ablation on EfficientNetV2B0 (FULL pipeline, standard fine-tuning), paired across the same 5 folds. The standard GAP head outperforms both attention and spatial-pooling variants ( $^\dagger$  denotes BH-FDR  $q < 0.05$ , paired  $t$ -test).

| Variant         | Standard head (%) | Variant (%) | $\Delta$ (pp) | Hedges' $g$ | $p$ ( $t$ )       |
|-----------------|-------------------|-------------|---------------|-------------|-------------------|
| attention       | 99.83             | 99.74       | +0.09         | +1.33       | 0.0205 $^\dagger$ |
| spatial_pooling | 99.83             | 96.15       | +3.68         | +1.44       | 0.0159 $^\dagger$ |

RANDBG. In both scenarios, the standard head and standard fine-tuning approaches yield the best results. Three of the five backbones achieve over 99.95% accuracy for every class. Only MobileNetV3Small exhibits notable off-diagonal confusion, misclassifying paper and scissors by 1.1% to 2.4%. The **none** class demonstrates near-perfect performance across all five backbones, which aligns with the anticipated outcome of the hard distractor strategy described in §3.2. The classical landmark baseline does not yield per-class confusion data for the same internal split. Its overall performance metrics are reported separately in Table 11 and visualized alongside the CNN results in Figures 7 and 8.

Figure 5 presents a summary of per-class behavior for the five backbones, based on the aggregated confusion matrices in the FULL/standard setting. DenseNet121, ResNet50, and VGG16 each achieve over 99.7% accuracy in all classes. In contrast,

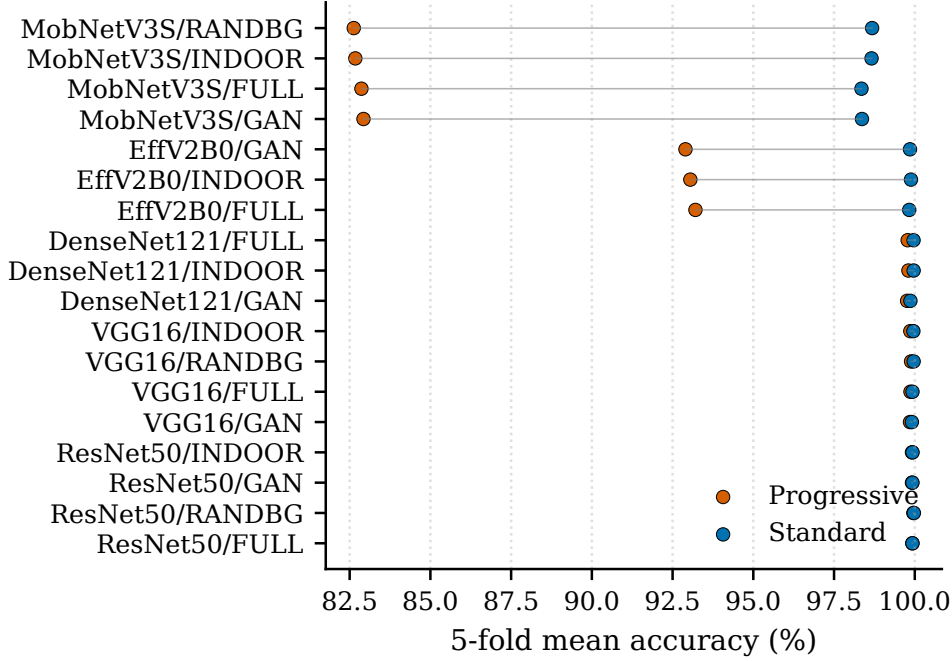


**Fig. 3** Paired comparison of standard and progressive fine-tuning for each backbone and ablation setting. Each row is one configuration; lines link the two protocols on the same five folds. The classical MediaPipe plus SVM baseline has no fine-tuning analogue and is therefore not shown here; see Figure 7 for its standalone accuracy.

MobileNetV3Small consistently exhibits a minor but reproducible confusion between the rock and paper classes (1% to 2.6%), accounting for the observed macro accuracy gap in Table 3.

In addition to the row-normalized view presented in Figure 5, Table 8 presents the per-class macro  $F_1$  mean and standard deviation across five folds for each backbone, head, and tuning configuration within the FULL pipeline. This approach reveals within-class variance that is not captured by the macro accuracy summary.

#### 4.6 Generalisation to HaGRID

A summary of HaGRID accuracy for internal versus direct transfer is presented for the FULL/standard cell in Table 9, with results visualized across all configurations in Figure 6. Although internal accuracy reaches ceiling levels ( $\geq 99\%$  on most cells), direct transfer accuracy varies from 66.2% to 87.7% across the 67 synthetic augmented configurations (mean gap  $-17.6$  pp, range  $[-33.7, -5.5]$  pp; NO\_SYNTH rows excluded). The largest residual gap occurs for VGG16 under progressive fine-tuning ( $-33.7$  percentage points), indicating that even extensive internal training augmentation does not fully address domain shift. Comprehensive per-cell results are presented in Table 10.

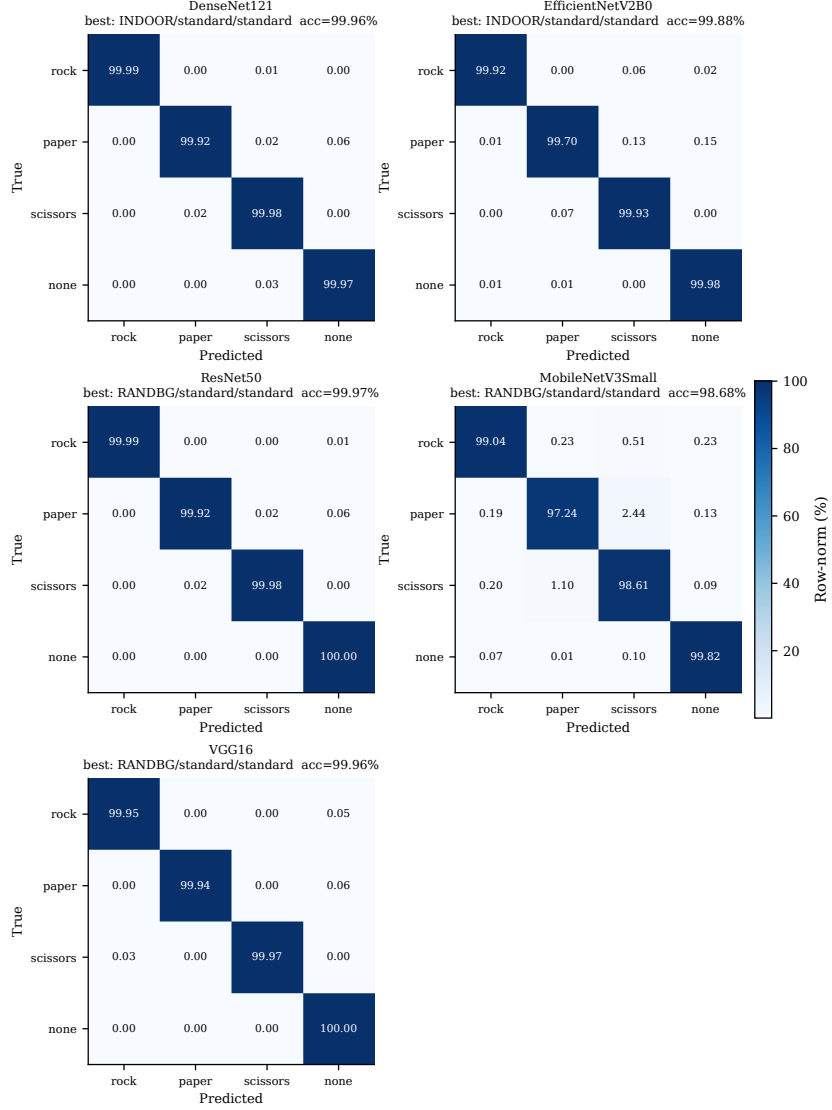


**Fig. 4** Confusion matrices for the best internal cell of each backbone, shown as row normalised percentages with two decimal places and aggregated over all five grouped image folds. Each panel uses the best complete fold cell for that backbone; the panel title gives the winning ablation, head and fine-tuning combination together with the resulting mean accuracy.

## 4.7 Classical Landmark Baseline

Table 11 presents the results for the auxiliary MediaPipe Hands combined with SVM. The classifier achieves 59.17% accuracy and 58.04% macro  $F_1$ . These outcomes are evaluated on the same four-class external subset used for direct transfer assessment.

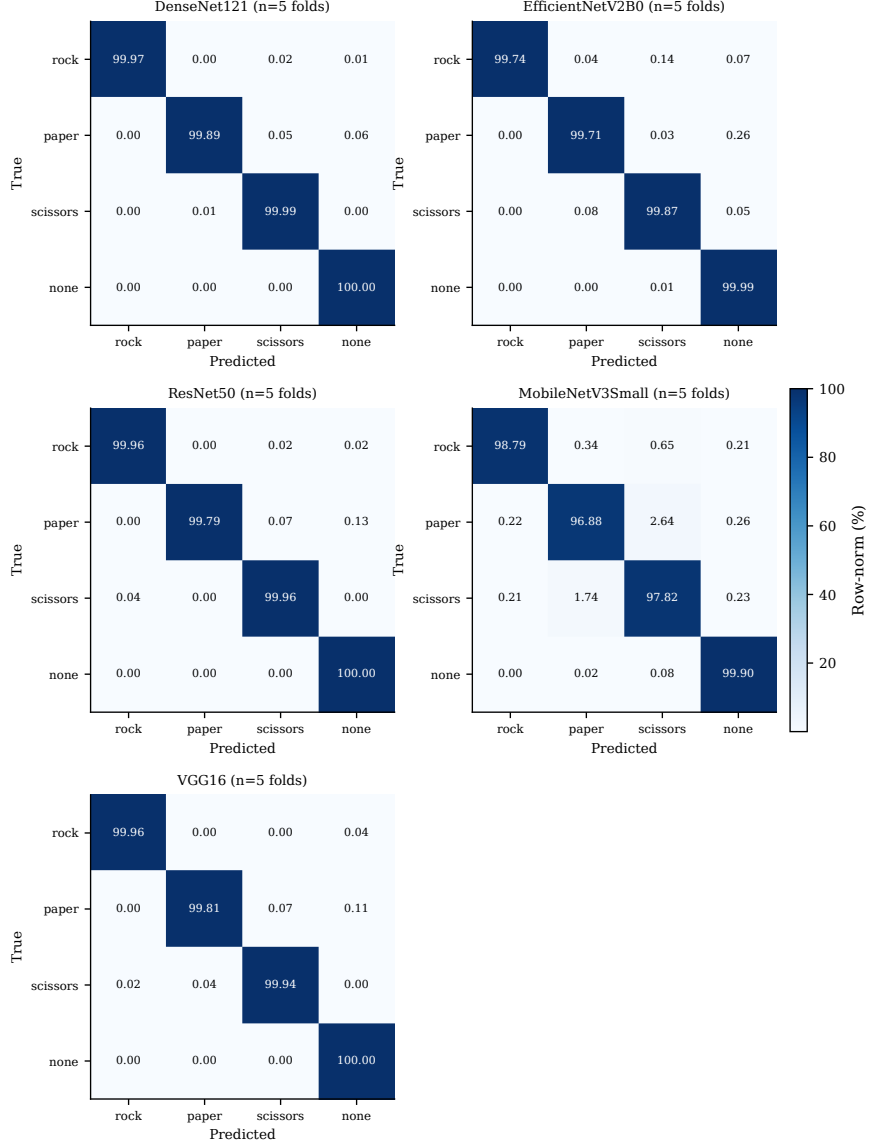


**Fig. 5** Row normalised confusion matrices in percent (two decimals), aggregated over all five grouped image folds, for the five FULL/standard head/standard fine-tuning configurations (DenseNet121, EfficientNetV2B0, ResNet50, MobileNetV3Small, VGG16). Rows are the true class and columns the predicted class. The heavy backbones saturate every diagonal cell above 99.7%; the lightweight MobileNetV3Small alone shows a measurable rock $\leftrightarrow$ paper confusion of 1 to 2.6%, consistent with its limited capacity. The classical MediaPipe plus SVM baseline does not produce per fold confusion matrices on this internal split, so it is reported separately in Table 11 and contrasted with the CNN backbones in Figures 7 and 8.

The primary limitation is not the SVM training time (0.38s following feature extraction), but rather the availability of landmarks. MediaPipe identifies a hand in

**Table 8** Internal per-class  $F_1$  (%) stability across the five image-grouped folds for every backbone/head/tune cell on the FULL pipeline (mean $\pm$ SD).

| Backbone         | Head            | Tune        | Rock $F_1$       | Paper $F_1$      | Scissors $F_1$   | None $F_1$       |
|------------------|-----------------|-------------|------------------|------------------|------------------|------------------|
| DenseNet121      | Standard        | Progressive | 99.91 $\pm$ 0.03 | 99.65 $\pm$ 0.04 | 99.73 $\pm$ 0.04 | 99.84 $\pm$ 0.01 |
| DenseNet121      | Standard        | Standard    | 99.98 $\pm$ 0.01 | 99.94 $\pm$ 0.07 | 99.96 $\pm$ 0.03 | 99.96 $\pm$ 0.06 |
| EfficientNetV2B0 | Standard        | Progressive | 94.08 $\pm$ 0.39 | 89.87 $\pm$ 0.50 | 90.57 $\pm$ 0.52 | 98.25 $\pm$ 0.04 |
| EfficientNetV2B0 | Attention       | Progressive | 94.30 $\pm$ 0.70 | 89.83 $\pm$ 0.90 | 90.39 $\pm$ 0.65 | 98.75 $\pm$ 0.13 |
| EfficientNetV2B0 | Attention       | Standard    | 99.83 $\pm$ 0.07 | 99.63 $\pm$ 0.04 | 99.67 $\pm$ 0.05 | 99.84 $\pm$ 0.03 |
| EfficientNetV2B0 | Standard        | Standard    | 99.87 $\pm$ 0.07 | 99.79 $\pm$ 0.04 | 99.84 $\pm$ 0.08 | 99.81 $\pm$ 0.02 |
| EfficientNetV2B0 | Spatial Pooling | Unknown     | 97.12 $\pm$ 1.52 | 94.09 $\pm$ 3.19 | 94.32 $\pm$ 3.06 | 99.04 $\pm$ 0.45 |
| MobileNetV3Small | Standard        | Progressive | 82.97 $\pm$ 0.60 | 75.79 $\pm$ 1.70 | 76.97 $\pm$ 0.90 | 95.24 $\pm$ 0.49 |
| MobileNetV3Small | Standard        | Standard    | 99.18 $\pm$ 0.06 | 97.37 $\pm$ 0.13 | 97.25 $\pm$ 0.12 | 99.60 $\pm$ 0.06 |
| ResNet50         | Standard        | Progressive | 99.97 $\pm$ 0.03 | 99.87 $\pm$ 0.01 | 99.98 $\pm$ 0.02 | 99.87 $\pm$ 0.02 |
| ResNet50         | Standard        | Standard    | 99.96 $\pm$ 0.04 | 99.90 $\pm$ 0.10 | 99.93 $\pm$ 0.06 | 99.92 $\pm$ 0.07 |
| VGG16            | Standard        | Progressive | 99.93 $\pm$ 0.05 | 99.80 $\pm$ 0.02 | 99.93 $\pm$ 0.02 | 99.80 $\pm$ 0.05 |
| VGG16            | Standard        | Standard    | 99.97 $\pm$ 0.05 | 99.89 $\pm$ 0.13 | 99.93 $\pm$ 0.09 | 99.92 $\pm$ 0.07 |

**Table 9** Internal vs. external (HaGRID zero-shot) accuracy on the FULL pipeline, standard head. A negative gap indicates a large drop on the unseen domain and is the central reviewer-relevant signal of synthetic-to-real distribution shift. Full results across every (backbone, ablation) cell appear in Table 10.

| Backbone         | Tune     | In-dist. (%) | Zero-shot (%) | Gap (pp) |
|------------------|----------|--------------|---------------|----------|
| DenseNet121      | Standard | 99.96        | 86.33         | -13.63   |
| ResNet50         | Standard | 99.93        | 83.83         | -16.10   |
| EfficientNetV2B0 | Standard | 99.83        | 80.17         | -19.66   |
| MobileNetV3Small | Standard | 98.35        | 80.17         | -18.18   |
| VGG16            | Standard | 99.93        | 79.83         | -20.10   |

72.97% of the training images, but only in 37.33% of the external images. This discrepancy accounts for the baseline error profile. The **none** class is accurately recovered from the zero vector, whereas gesture classes exhibit reduced performance when the hand detector fails under the external distribution.

Consequently, this result serves as a useful auxiliary comparator. However, it is not included in the paired CNN ablation experiments described in §3.11.

Figure 7 compares the baseline with the five CNN backbones along a unified accuracy axis.

The weakest CNN configuration on HaGRID under direct transfer (approximately 80%) exceeds the MediaPipe plus SVM external accuracy (59.2%) by more than 20 percentage points. This result demonstrates that the synthetic augmentation combined with the CNN pipeline maintains a substantial generalization margin over the landmark-based baseline, even when the CNN experiences a reduction of at least 13 percentage points due to the synthetic-to-real gap.

**Table 10** Best (by HaGRID zero-shot accuracy) head/tune setting for every (backbone, ablation) pair: in-distribution accuracy, zero-shot accuracy, and generalisation gap (% , percentage points). Sorted alphabetically by backbone then ablation.

| Backbone         | Ablation       | Head      | Tune        | In-dist. (%) | Zero-shot (%) | Gap (pp) |
|------------------|----------------|-----------|-------------|--------------|---------------|----------|
| DenseNet121      | FULL           | Standard  | Standard    | 99.96        | 86.33         | -13.63   |
| DenseNet121      | GAN            | Standard  | Standard    | 99.86        | 85.83         | -14.03   |
| DenseNet121      | INDOOR         | Standard  | Progressive | 99.80        | 84.67         | -15.14   |
| DenseNet121      | NO_ALPHA       | Standard  | Standard    | 99.80        | 85.67         | -14.13   |
| DenseNet121      | NO_SHIFT       | Standard  | Standard    | 99.73        | 87.33         | -12.39   |
| DenseNet121      | NO_SYNTH       | Standard  | Standard    | 56.70        | 64.17         | +7.47    |
| DenseNet121      | RANDBG         | Standard  | Progressive | 99.82        | 86.83         | -12.98   |
| DenseNet121      | RATIO_1X       | Standard  | Standard    | 99.77        | 84.67         | -15.10   |
| DenseNet121      | REMBG_ONLY     | Standard  | Standard    | 99.86        | 82.00         | -17.86   |
| DenseNet121      | STYLE_TRANSFER | Standard  | Standard    | 99.66        | 87.67         | -12.00   |
| EfficientNetV2B0 | FULL           | Attention | Standard    | 99.74        | 80.17         | -19.58   |
| EfficientNetV2B0 | GAN            | Standard  | Standard    | 99.85        | 83.50         | -16.35   |
| EfficientNetV2B0 | INDOOR         | Standard  | Standard    | 99.88        | 79.33         | -20.55   |
| EfficientNetV2B0 | NO_ALPHA       | Standard  | Standard    | 98.25        | 81.00         | -17.25   |
| EfficientNetV2B0 | NO_SHIFT       | Standard  | Standard    | 95.20        | 81.00         | -14.20   |
| EfficientNetV2B0 | NO_SYNTH       | Standard  | Standard    | 43.10        | 63.50         | +20.40   |
| EfficientNetV2B0 | RANDBG         | Standard  | Standard    | 99.87        | 79.67         | -20.20   |
| EfficientNetV2B0 | RATIO_1X       | Standard  | Standard    | 98.26        | 79.00         | -19.26   |
| EfficientNetV2B0 | REMBG_ONLY     | Standard  | Standard    | 94.72        | 79.67         | -15.06   |
| EfficientNetV2B0 | STYLE_TRANSFER | Standard  | Standard    | 95.42        | 82.00         | -13.42   |
| MobileNetV3Small | FULL           | Standard  | Standard    | 98.35        | 80.17         | -18.18   |
| MobileNetV3Small | GAN            | Standard  | Standard    | 98.36        | 84.00         | -14.36   |
| MobileNetV3Small | INDOOR         | Standard  | Standard    | 98.66        | 82.33         | -16.33   |
| MobileNetV3Small | NO_ALPHA       | Standard  | Standard    | 91.56        | 81.50         | -10.06   |
| MobileNetV3Small | NO_SHIFT       | Standard  | Standard    | 88.47        | 81.00         | -7.47    |
| MobileNetV3Small | NO_SYNTH       | Standard  | Standard    | 41.81        | 58.00         | +16.19   |
| MobileNetV3Small | RANDBG         | Standard  | Standard    | 98.68        | 81.83         | -16.85   |
| MobileNetV3Small | RATIO_1X       | Standard  | Standard    | 88.48        | 83.00         | -5.48    |
| MobileNetV3Small | REMBG_ONLY     | Standard  | Standard    | 94.85        | 82.17         | -12.69   |
| MobileNetV3Small | STYLE_TRANSFER | Standard  | Standard    | 91.64        | 81.00         | -10.64   |
| ResNet50         | FULL           | Standard  | Standard    | 99.93        | 83.83         | -16.10   |
| ResNet50         | GAN            | Standard  | Standard    | 99.93        | 84.50         | -15.43   |
| ResNet50         | INDOOR         | Standard  | Standard    | 99.93        | 84.33         | -15.60   |
| ResNet50         | NO_ALPHA       | Standard  | Standard    | 99.80        | 82.83         | -16.96   |
| ResNet50         | NO_SHIFT       | Standard  | Standard    | 99.85        | 86.50         | -13.35   |
| ResNet50         | NO_SYNTH       | Standard  | Standard    | 52.79        | 62.17         | +9.38    |
| ResNet50         | RANDBG         | Standard  | Standard    | 99.97        | 85.83         | -14.14   |
| ResNet50         | RATIO_1X       | Standard  | Standard    | 99.81        | 85.33         | -14.48   |
| ResNet50         | REMBG_ONLY     | Standard  | Standard    | 99.87        | 83.50         | -16.37   |
| ResNet50         | STYLE_TRANSFER | Standard  | Standard    | 99.81        | 87.33         | -12.48   |
| VGG16            | FULL           | Standard  | Standard    | 99.93        | 79.83         | -20.10   |
| VGG16            | GAN            | Standard  | Standard    | 99.91        | 85.67         | -14.24   |
| VGG16            | INDOOR         | Standard  | Standard    | 99.96        | 75.83         | -24.12   |
| VGG16            | NO_ALPHA       | Standard  | Standard    | 99.84        | 76.83         | -23.01   |
| VGG16            | NO_SHIFT       | Standard  | Standard    | 99.82        | 76.67         | -23.15   |
| VGG16            | NO_SYNTH       | Standard  | Standard    | 45.70        | 55.33         | +9.64    |
| VGG16            | RANDBG         | Standard  | Standard    | 99.96        | 75.00         | -24.96   |
| VGG16            | RATIO_1X       | Standard  | Standard    | 99.82        | 75.00         | -24.82   |
| VGG16            | REMBG_ONLY     | Standard  | Standard    | 99.84        | 72.00         | -27.84   |
| VGG16            | STYLE_TRANSFER | Standard  | Standard    | 99.80        | 82.17         | -17.64   |

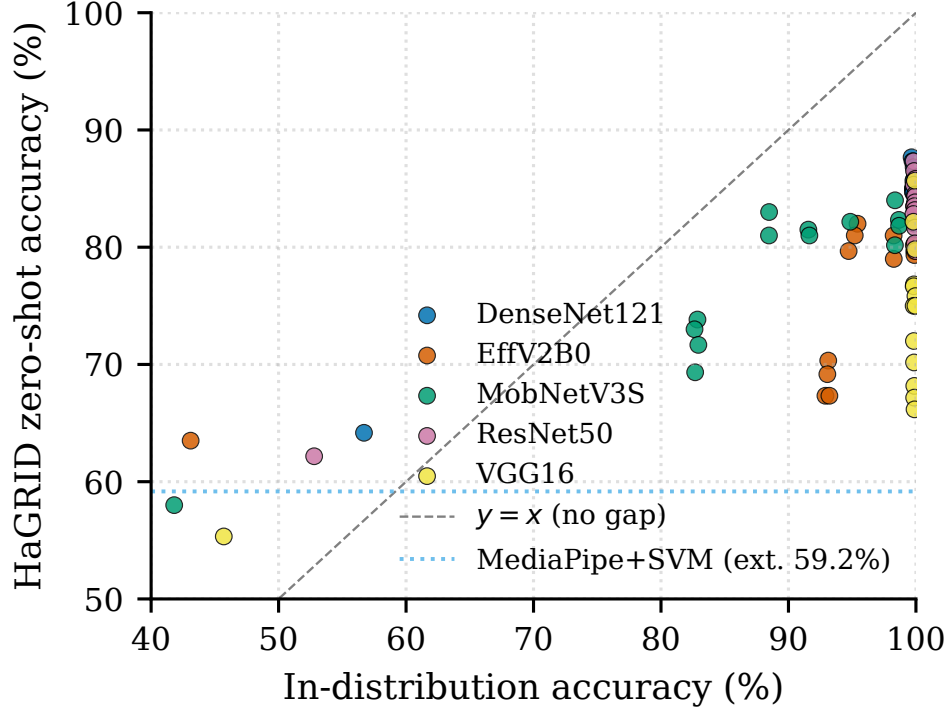


**Fig. 6** Internal versus direct transfer HaGRID accuracy across all evaluated configurations, coloured by backbone. The dashed  $y=x$  line marks parity between the two measurements.

**Table 11** Auxiliary MediaPipe Hands + RBF-SVM baseline on the retained four-class external slice. Training uses 7,830 raw RPS gesture images plus 2,610 seed-42 sampled background images for **none**; testing uses 600 external images (150 per class).

| Metric                          | Value  |
|---------------------------------|--------|
| Training samples                | 10,440 |
| External test samples           | 600    |
| External accuracy (%)           | 59.17  |
| External macro- $F_1$ (%)       | 58.04  |
| Training landmark detection (%) | 72.97  |
| External landmark detection (%) | 37.33  |
| SVM training time (s)           | 0.38   |

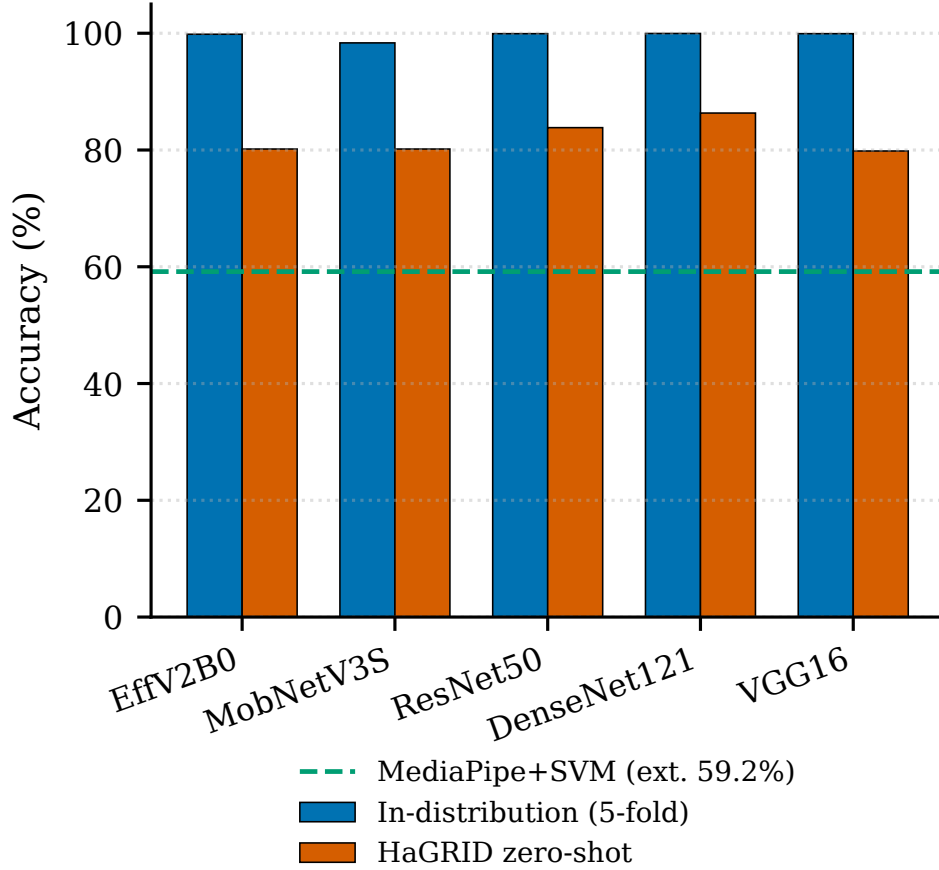


**Fig. 7** Comparison across modelling paradigms: internal fivefold validation accuracy for each backbone (blue) and HaGRID direct transfer accuracy (orange, best head/tune per backbone in the FULL pipeline) plotted against the auxiliary MediaPipe Hands plus SVM external accuracy (green dashed reference, 59.2%, T14).

#### 4.8 Per Class Behavior under Direct Transfer

Aggregated macro accuracy obscures the distribution of errors across individual classes following transfer. Table 12 presents the five strongest and five weakest configurations based on per-class direct transfer  $F_1$ . This analysis highlights which model designs prioritize gesture class performance at the expense of the **none**/background class.

Table 13 summarizes the tradeoff between macro precision and macro recall for the same experimental setting. Rows exhibiting the largest absolute gap indicate operating points where a single deployment threshold would result in poor calibration.

These complementary perspectives demonstrate that direct transfer macro accuracy may obscure significant differences in per-class error profiles across various

**Table 12** Per-class HaGRID zero-shot  $F_1$  (mean over folds): top-5 and bottom-5 configurations by mean per-class  $F_1$  (out of N=104 unique backbone/ablation/head/tune cells). Configurations differ substantially in how they sacrifice the **none**/background class under domain shift.

| Backbone                                           | Ablation       | Tune        | Rock $F_1$ | Paper $F_1$ | Scissors $F_1$ | None $F_1$ | Mean $F_1$   |
|----------------------------------------------------|----------------|-------------|------------|-------------|----------------|------------|--------------|
| <i>Top-5 by zero-shot mean <math>F_1</math></i>    |                |             |            |             |                |            |              |
| DenseNet121                                        | STYLE_TRANSFER | Standard    | 0.879      | 0.918       | 0.899          | 0.836      | <b>0.883</b> |
| DenseNet121                                        | NO_ALPHA       | Standard    | 0.873      | 0.908       | 0.887          | 0.823      | <b>0.873</b> |
| DenseNet121                                        | NO_ALPHA       | Progressive | 0.837      | 0.927       | 0.912          | 0.813      | <b>0.872</b> |
| DenseNet121                                        | NO_SHIFT       | Standard    | 0.865      | 0.905       | 0.885          | 0.820      | <b>0.869</b> |
| DenseNet121                                        | GAN            | Progressive | 0.828      | 0.931       | 0.912          | 0.804      | <b>0.869</b> |
| <i>Bottom-5 by zero-shot mean <math>F_1</math></i> |                |             |            |             |                |            |              |
| MobileNetV3Small                                   | NO_SYNTH       | Progressive | 0.583      | 0.497       | 0.604          | 0.071      | <b>0.439</b> |
| VGG16                                              | NO_SYNTH       | Progressive | 0.566      | 0.738       | 0.478          | 0.002      | <b>0.446</b> |
| VGG16                                              | NO_SYNTH       | Standard    | 0.608      | 0.707       | 0.520          | 0.003      | <b>0.459</b> |
| DenseNet121                                        | NO_SYNTH       | Progressive | 0.746      | 0.643       | 0.489          | 0.000      | <b>0.470</b> |
| EfficientNetV2B0                                   | NO_SYNTH       | Progressive | 0.611      | 0.563       | 0.567          | 0.152      | <b>0.473</b> |

**Table 13** Macro precision vs. macro recall trade-off on the HaGRID zero-shot test (mean across folds, in %). Top section: best configurations by macro  $F_1$ . Bottom section: configurations with the largest absolute precision–recall imbalance, indicating threshold-sensitive operating points (out of N=104 unique backbone/ablation/head/tune cells).

| Backbone                                      | Ablation       | Tune        | Macro P | Macro R | P–R gap | Macro $F_1$ |
|-----------------------------------------------|----------------|-------------|---------|---------|---------|-------------|
| <i>Top-5 by macro <math>F_1</math></i>        |                |             |         |         |         |             |
| DenseNet121                                   | STYLE_TRANSFER | Standard    | 90.51   | 88.00   | +2.51   | 89.24       |
| DenseNet121                                   | NO_ALPHA       | Progressive | 90.33   | 86.90   | +3.43   | 88.58       |
| DenseNet121                                   | GAN            | Progressive | 90.38   | 86.50   | +3.88   | 88.40       |
| DenseNet121                                   | NO_ALPHA       | Standard    | 89.85   | 86.94   | +2.91   | 88.37       |
| DenseNet121                                   | STYLE_TRANSFER | Progressive | 90.11   | 86.54   | +3.57   | 88.29       |
| <i>Top-5 by absolute precision–recall gap</i> |                |             |         |         |         |             |
| VGG16                                         | INDOOR         | Progressive | 84.87   | 67.70   | +17.16  | 75.32       |
| VGG16                                         | REMBG_ONLY     | Progressive | 82.21   | 65.90   | +16.31  | 73.16       |
| VGG16                                         | FULL           | Progressive | 84.39   | 68.27   | +16.12  | 75.48       |
| VGG16                                         | NO_ALPHA       | Progressive | 85.18   | 71.43   | +13.75  | 77.70       |
| VGG16                                         | NO_SHIFT       | Progressive | 85.64   | 71.92   | +13.72  | 78.18       |

configurations. Figure 8 provides a comprehensive backbone-specific analysis by presenting the per-class  $F_1$  scores of each backbone’s optimal direct transfer cell on the reserved four-class HaGRID subset, together with the corresponding macro  $F_1$  values. The auxiliary MediaPipe Hands combined with the SVM landmark baseline (T14) is included as a horizontal reference. All per-class CNN  $F_1$  scores exceed 76%, and all backbones except EfficientNetV2B0 maintain the **none** class above 80% despite domain shift. In contrast, the landmark baseline achieves only 58% macro  $F_1$  overall. This comparison situates the performance gap between paradigms on the same class

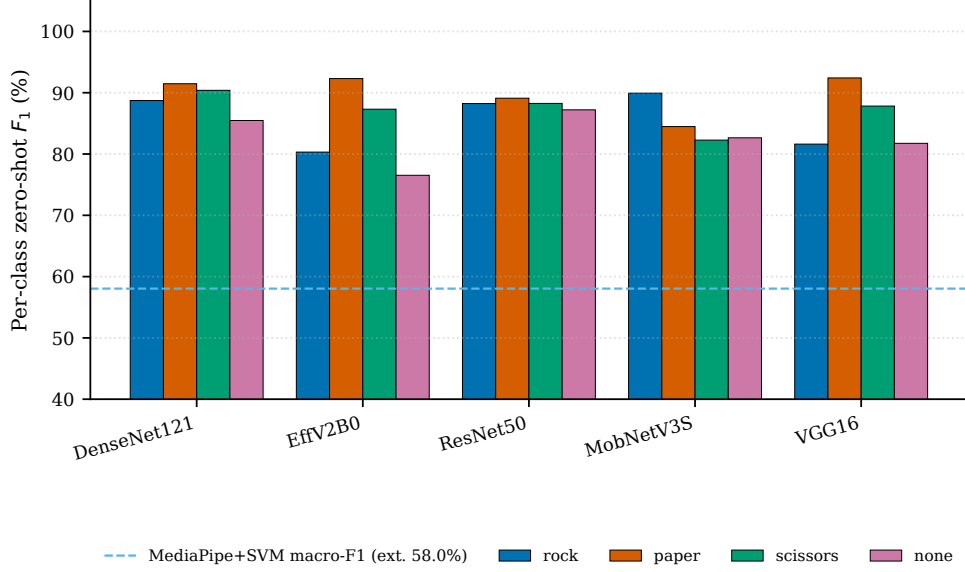


**Fig. 8** Per class direct transfer  $F_1$  (%) on the reserved four class external slice for each of the five CNN backbones (best cell per backbone); the dashed reference is the MediaPipe Hands plus SVM landmark baseline’s overall macro  $F_1$  (58.0%, T14). A per class breakdown is not available for the landmark baseline, hence its reference line is uniform across backbones.

axis as the CNN analysis. It demonstrates that the augmentation combined with the CNN pipeline outperforms the baseline both on average and within each class, even after accounting for the synthetic-to-real domain gap.

#### 4.9 Visual Evidence from Grad CAM

To determine whether the reported gains originate from gesture-related attention rather than background memorization, Grad CAM activation maps were extracted for all five backbones. A correctly classified sample (top row) was contrasted with a misclassified one (bottom row), resulting in the side-by-side panel of Figure 9. In the correct row, the activation mass consistently concentrates on the hand region for every backbone, even in visually complex composite backgrounds such as clothing racks, train stations, and indoor living room scenes. This observation supports the claim that the internal accuracy reported in Tables 3 and 9 is primarily driven by gesture-related errors. The error row reveals a complementary pattern: each backbone exhibits distinct failure characteristics. DenseNet121 (under the **RATIO\_1X** ablation) divides its attention between the hand and a salient bystander in the background, resulting in the misclassification of *rock* as *scissors*. EfficientNetV2B0 and ResNet50 maintain focus on the hand but misinterpret the finger geometry as *scissors*; notably, ResNet50 does so with 99% confidence, indicating a calibration failure mode that should be considered for downstream deployment.

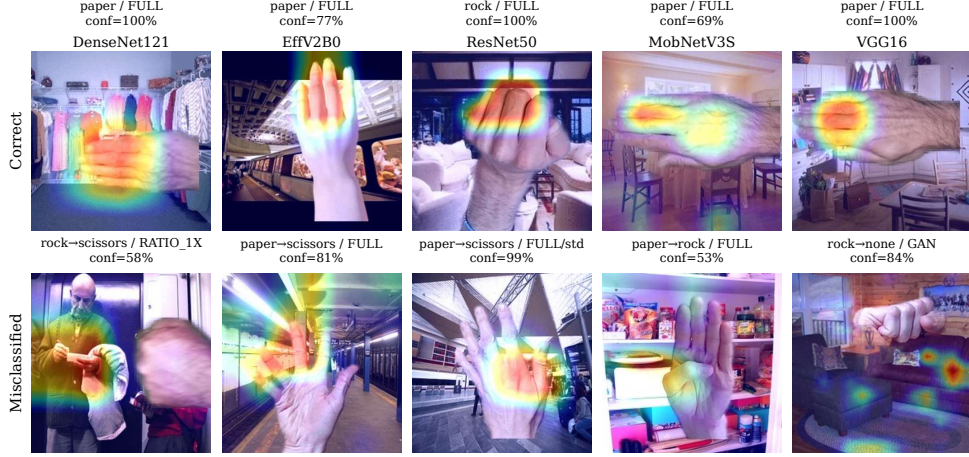


**Fig. 9** Grad CAM activation maps for each backbone. Top row: a correctly classified sample for each of the five backbones (FULL pipeline, progressive fine-tuning where available). Bottom row: a misclassified sample sourced from the most informative ablation cell that produced one (RATIO\_1X for DenseNet121, GAN for VGG16, FULL otherwise; cell label reports true  $\rightarrow$  predicted class and configuration). Each panel is annotated with the prediction confidence; the error row reveals characteristic failure modes (background bystander, finger geometry confusion, attention loss, distractor shift). Grad CAM does not apply to the landmark based MediaPipe plus SVM baseline, so it is deliberately excluded from this figure; its per class behaviour is summarised in Figure 8 and Table 11.

MobileNetV3Small fails to maintain attention on the upright fingers and misclassifies *paper* as *rock* with near chance confidence, which aligns with its narrower receptive field and the rock $\leftrightarrow$ paper confusion observed in Figure 5. VGG16 (under GAN) directs its attention entirely away from the hand toward a bright background distractor, predicting *none* for a clear *rock* sample. The contrast between the two rows thus corroborates both the robustness of the FULL pipeline in most cases and the specific failure modes quantified by the backbone-specific confusion matrices in Figure 5.

#### 4.10 Temporal Stability for Deployment

Table 14 presents the optimal result for each window size in the temporal smoothing grid (§3.10) evaluated on the deployment EfficientNetV2B0/FULL stream.

The optimal configuration is a single frame ( $N=1$ ) with a low threshold ( $\theta=0.4$ ), achieving 94.17% accuracy and a 0% false positive rate on the **none** class. Longer windows ( $N \geq 3$ ) become competitive only when  $\theta$  approaches 0.9, at which point per-class accuracy declines to below 50%.

These results indicate that the trained classifier is well-calibrated on a frame-by-frame basis. Aggressive temporal smoothing reduces the useful signal in exchange for only marginal improvements in false-positive control in this context.

**Table 14** Best per-window result of the temporal-smoothing grid search on the deployment-side EfficientNetV2B0/FULL stream ( $N \in \{1, 3, 5, 7, 10, 15\}$ ,  $\theta \in [0.4, 0.9]$ , score = accuracy – false-positive rate). The optimum ( $\star$ ) is the single-frame, low-threshold setting; longer windows trade accuracy for false positives at this confidence floor.

| $N$ | $\theta$ | Accuracy (%) | FPR (%) | Score          |
|-----|----------|--------------|---------|----------------|
| 1   | 0.4      | 94.17        | 0.00    | 0.9417 $\star$ |
| 3   | 0.7      | 80.06        | 0.22    | 0.7972         |
| 5   | 0.8      | 66.17        | 0.89    | 0.6483         |
| 7   | 0.9      | 47.83        | 0.00    | 0.4783         |
| 10  | 0.9      | 40.61        | 0.22    | 0.4028         |
| 15  | 0.9      | 28.22        | 6.67    | 0.1822         |

## 5 Discussion

Taken together, the findings suggest five broader lessons for small RGB gesture datasets.

### *Synthetic augmentation carries most of the gain.*

The most prominent finding in this study is the significant influence of the data pipeline on performance. In 43 paired comparisons between a synthetic setting and the baseline, background compositing increases accuracy from approximately 48% to at least 96%. When the FULL pipeline is implemented, the five backbones exhibit much more similar performance during internal validation. For this task, the primary limitation is associated with the design of the training distribution rather than backbone selection. This observation is consistent with recent research on hybrid and generative AI augmentation [5, 6].

### *Progressive unfreezing does not help here.*

A paired comparison between standard and progressive fine-tuning yields a notable negative result: discriminative learning rate schedules impair EfficientNetV2B0 and MobileNetV3Small performance rather than improving it. This finding aligns with the broader fine-tuning literature [9, 10], although such negative outcomes are seldom reported explicitly in static gesture studies. The prevailing interpretation is that the combination of lightweight backbones and frozen Batch Normalization statistics provides limited opportunity for progressive unfreezing to be beneficial. However, a more mechanistic analysis is required to substantiate this explanation.

### *The domain gap remains substantial.*

The mean  $-17.6$ pp gap between internal validation and HaGRID, with individual configurations ranging from  $-5.5$  to  $-33.7$ pp across the 67 synthetic augmented cells, represents the most practically significant finding and an important result in the paper.

The pipeline demonstrates an improvement over NO.SYNTH by more than 40 pp; however, the persistent performance gap indicates that photorealistic compositing does not yet match the real-world distribution. Bridging this gap may necessitate richer compositing techniques, increased background diversity, or more adaptive approaches to domain randomization.

***The landmark baseline is limited by detection recall.***

The MediaPipe combined with the support vector machine (SVM) baseline fails for a reason distinct from that of the convolutional neural network (CNN) grid. After landmark detection, a shallow classifier is able to separate a substantial portion of the samples. However, the primary limitation arises from hand detection performance on the external slice, which decreases to 37.3%. These missed detections result in zero vector collisions, which cause gesture examples to be misclassified as belonging to the `none` class. In this rock-paper-scissors (RPS) context, pixel-based CNNs trained with background randomization maintain a significantly stronger margin under distributional shift compared to the landmark-based baseline.

***Evaluation protocol changes the story the scores tell.***

When considered collectively, the grouped cross-validation results and the reserved HaGRID evaluation underscore a critical point: near-ceiling internal accuracy alone does not constitute evidence of robust generalization. For this reason, the methodological contribution is as significant as the model ablation. The protocol maintains the visibility of the residual gap rather than allowing it to be obscured by a superficially clean but forgiving data split.

## 6 Limitations

***Subject grouping.***

The Kaggle slice used here does not expose subject identifiers. Our cross validation protocol blocks augmentation leakage, but it cannot fully rule out subject leakage. Replication on the full HaGRID release with its `user_id` field would tighten this guarantee. The reserved HaGRID evaluation partly constrains the risk: if subject overlap were the main driver of internal performance, one would expect a much sharper collapse on disjoint subjects than the observed best case of 87.7%. This does not remove the concern, but it suggests that any remaining subject leakage is unlikely to explain more than a modest share of the gain.

***Statistical precision and seeding.***

When  $n = 5$  folds and a single random seed are used, inferential precision is inherently limited. The minimum achievable  $p$ -value for a paired Wilcoxon signed-rank test at this sample size, determined by the number of pairwise comparisons, is approximately 0.0625. Employing multiple random seeds (for example,  $3 \times 5 = 15$  paired observations per cell) increases the number of ablation scenarios, enabling a more granular assessment. More broadly, the standard deviations presented in the tables indicate variation across folds within a single split realization. These values do not capture estimator

variance across independent repeated splits, which the use of multiple seeds would also help to quantify.

***Incomplete fold for the headline grid cell.***

The DenseNet121/RANDBG/standard cell achieves the highest mean accuracy in the full grid; however, results are reported for  $n = 4$  folds due to the unavailability of fold 1 in the locally retained artifact (Table A1). Accordingly, Figure 4 presents the best complete fold cell for each backbone individually; thus, DenseNet121 is represented by its INDOOR/std/std cell (99.96%,  $n = 5$ ). The headline estimate for DenseNet121 retains the  $n = 4$  limitation; however, the conclusions in §4 remain unaffected, as the 43 synthetic versus baseline contrasts are calculated exclusively on cells with complete fold coverage.

***Gap between synthetic training and real world transfer.***

Direct transfer accuracy on HaGRID ranges from 66.2% to 87.7% across the 67 synthetic augmented configurations (mean gap  $-17.6$  pp; NO\_SYNTH excluded), indicating that synthetic compositing alone does not fully bridge the real world distribution.

***Comparability of the classical baseline.***

The MediaPipe combined with SVM serves as an auxiliary external slice diagnostic rather than a 73-cell equivalent to the CNN grid. The publicly available code provides a `baseline_mediapipe_svm.py` implementation for a split `synthetic_full` directory; however, the retained local artifacts do not include this split or its corresponding saved report. Accordingly, the reproducible run trained on raw gesture images and background negatives, and evaluated on the external slice, is reported separately and excluded from the paired statistical tests.

***Composite artefacts.***

The segmentation probe in §3.2 provides only a proxy IoU sanity check against MediaPipe hand boxes; it is not a Dice or IoU score against pixel level human masks. Some hard edges, imperfect finger boundaries, and foreground or background lighting mismatches may remain in the generated images.

***Temporal smoothing.***

A retained deployment diagnostic for the EfficientNetV2B0/FULL cell searched smoothing windows  $N \in \{1, 3, 5, 7, 10, 15\}$  and decision thresholds  $\{0.4, \dots, 0.9\}$ . The best score was obtained at  $N = 1$ , threshold 0.4 (score 94.17%, false positive rate 0), so no temporal smoothing is used in the static image results reported in this paper.

***Class scope.***

Only four classes (rock, paper, scissors, none) are evaluated; generalisation to the broader HaGRID 18-class taxonomy is not addressed.

### *Real time deployment.*

Inference latency on edge hardware is reported only at the data generation stage; full deployment cost (mixed precision inference plus the **rembg** preprocessing) is not measured.

## 7 Conclusion

The results demonstrate that synthetic scene augmentation substantially enhances static hand gesture recognition, provided that evaluation protocols prevent image-level leakage and ensure fair comparisons across backbones. Under fivefold evaluation with image-level grouping, the proposed pipeline increases accuracy from approximately 48% to at least 96% across five convolutional neural network (CNN) backbones. These improvements remain statistically significant after correction for multiple comparisons. Two key findings support this conclusion. First, progressive fine-tuning negatively affects lightweight backbones in this context. Second, a mean  $-17.6$  pp performance gap persists on HaGRID across the 67 synthetic augmentation configurations. The broader implication is clear: for small RGB gesture datasets, rigorous evaluation protocols and substantial background variation are more critical than increasingly complex model heads or tuning schedules.

Code, raw per-fold outputs, and reproduction scripts will be released under an open license at <https://github.com/RageLog/bg-robust-rps>.

**Acknowledgements.** The authors thank the maintainers of the public datasets used in this study.

## Declarations

**Funding** No external funding was received for this work.

**Competing interests** The authors declare that they have no competing interests.

**Ethics approval and consent to participate** Not applicable; the study uses only publicly available, anonymised image datasets.

**Data availability** All datasets analysed in this study are publicly available from the sources cited in Section 3.1.

**Code availability** Code, raw per-fold JSON outputs, and reproduction scripts are available at <https://github.com/RageLog/bg-robust-rps>.

**Author contributions** C.T. and R.A. designed the study and the experimental protocol. C.T. implemented the pipeline, conducted the experiments, and prepared the figures and tables. R.A. supervised the work and contributed to the statistical analysis. Both authors wrote and approved the final manuscript.

## References

- [1] Mohanty, A., Roy, K., Sahay, R.R.: Robust static hand gesture recognition: Harnessing sparsity of deeply learned features. *The Visual Computer* **40**(9), 6507–6531 (2024) <https://doi.org/10.1007/s00371-023-03179-0>

- [2] Niu, P.: Convolutional neural network for gesture recognition human-computer interaction system design. *PLoS ONE* **20**(2), 0311941 (2025) <https://doi.org/10.1371/journal.pone.0311941>
- [3] Darmawan, I., Meldiawati, A.M., Rizal, R., Rahmatulloh, A., Haerani, E., Gunawan, R.: Hands-On gaming: Real-time gesture detection in rock-paper-scissors using Mediapipe and CNN. In: *Proc. Int. Conf. Adv. Data Sci. E-Learn. Inf. Syst. (ICADEIS)* (2025). <https://doi.org/10.1109/ICADEIS65852.2025.10933061>
- [4] Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P.: Domain randomization for transferring deep neural networks from simulation to the real world. In: *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, pp. 23–30 (2017). <https://doi.org/10.1109/IROS.2017.8202133>
- [5] Awaluddin, B.-A., Chao, C.-T., Chiou, J.-S.: A hybrid image augmentation technique for user- and environment-independent hand gesture recognition based on deep learning. *Mathematics* **12**(9), 1393 (2024) <https://doi.org/10.3390/math12091393>
- [6] Rahat, F., Hossain, M..S., Ahmed, M.R., Jha, S.K., Ewetz, R.: Data augmentation for image classification using generative AI. In: *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, pp. 4173–4182 (2025). <https://doi.org/10.1109/WACV61041.2025.00410>
- [7] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 770–778 (2016). <https://doi.org/10.1109/CVPR.2016.90>
- [8] Tan, M., Le, Q.V.: EfficientNetV2: Smaller models and faster training. In: *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, pp. 10096–10106 (2021). <http://proceedings.mlr.press/v139/tan21a.html>
- [9] Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 328–339. Association for Computational Linguistics, Melbourne, Australia (2018). <https://doi.org/10.18653/v1/P18-1031>
- [10] Smith, L.N.: Cyclical learning rates for training neural networks. In: *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 464–472 (2017). <https://doi.org/10.1109/WACV.2017.58>
- [11] Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. *Journal of Big Data* **6**(1), 60 (2019) <https://doi.org/10.1186/s40537-019-0197-0>

- [12] Cui, C., Sunar, M.S., Su, G.E.: Deep vision-based real-time hand gesture recognition: A review. *PeerJ Computer Science* **11**, 2921 (2025) <https://doi.org/10.7717/peerj-cs.2921>
- [13] DrGFreeman: Rock Paper Scissors Dataset. Kaggle. <https://www.kaggle.com/datasets/drgfreeman/rockpaperscissors> (2018)
- [14] Frtgmn: Rock Paper Scissors Dataset. Kaggle. <https://www.kaggle.com/datasets/frtgmn/rock-paper-scissor> (2020)
- [15] Sanikamal: Rock Paper Scissors Dataset. Kaggle. <https://www.kaggle.com/datasets/sanikamal/rockpaper-scissors-dataset> (2019)
- [16] Alexandredj: Rock Paper Scissors Dataset. Kaggle. <https://www.kaggle.com/datasets/alexandredj/rockpaper-scissors-rps> (2021)
- [17] Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 413–420 (2009). <https://doi.org/10.1109/CVPR.2009.5206848> . Dataset: <https://www.kaggle.com/datasets/itsahmad/indoor-scenes-cvpr-2019>
- [18] Nuzhdin, A., Nagaev, A., Sautin, A., Kapitanov, A., Kvanchiani, K.: HaGRIDv2: 1M Images for Static and Dynamic Hand Gesture Recognition (2024). <https://arxiv.org/abs/2412.01508>
- [19] innominate817: HaGRID Sample 30k 384p. Kaggle Dataset. <https://www.kaggle.com/datasets/innominate817/hagrid-sample-30k-384p> (2023)
- [20] Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O.R., Jagersand, M.: U<sup>2</sup>-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognition* **106**, 107404 (2020) <https://doi.org/10.1016/j.patcog.2020.107404>
- [21] Shrivastava, A., Gupta, A., Girshick, R.: Training region-based object detectors with online hard example mining. In: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 761–769 (2016). <https://doi.org/10.1109/CVPR.2016.89>
- [22] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 248–255 (2009). <https://doi.org/10.1109/CVPR.2009.5206848>
- [23] Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q.V., Adam, H.: Searching for MobileNetV3. In: *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 1314–1324 (2019). <https://doi.org/10.1109/ICCV.2019.00140>

- [24] Huang, G., Liu, Z., Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 2261–2269 (2017). <https://doi.org/10.1109/CVPR.2017.243>
- [25] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proc. 3rd Int. Conf. Learn. Represent. (ICLR) (2015). <http://arxiv.org/abs/1409.1556>
- [26] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proc. 3rd Int. Conf. Learn. Represent. (ICLR) (2015). <http://arxiv.org/abs/1412.6980>
- [27] Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., Wu, H.: Mixed precision training. In: Proc. 6th Int. Conf. Learn. Represent. (ICLR) (2018). <http://arxiv.org/abs/1710.03740>
- [28] Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M.G., Lee, J., Chang, W.-T., Hua, W., Georg, M., Grundmann, M.: MediaPipe: A Framework for Building Perception Pipelines (2019). <https://arxiv.org/abs/1906.08172>
- [29] Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**(3), 273–297 (1995) <https://doi.org/10.1007/BF00994018>
- [30] Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological) **57**(1), 289–300 (1995) <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>

## Appendix A Per-Configuration Cross Validation Results

Table A1 reports the validation accuracy and macro  $F_1$  for every one of the 73 evaluated configurations (fivefold cross validation with image level grouping, mean $\pm$ SD in %). This extended table is provided as a supplementary reference for readers seeking the raw cell level numbers behind the aggregated figures and tables of §4.

**Table A1** Five-fold cross-validation accuracy and macro- $F_1$  for all 73 evaluated configurations (mean $\pm$ SD, %). Rows are sorted by mean accuracy and reported in two consecutive blocks for readability.

| Backbone         | Ablation       | Head     | Tune        | Folds | Acc. (%)         | Macro- $F_1$ (%) |
|------------------|----------------|----------|-------------|-------|------------------|------------------|
| DenseNet121      | RANDBG         | Standard | Standard    | 4     | 99.98 $\pm$ 0.02 | 99.98 $\pm$ 0.02 |
| ResNet50         | RANDBG         | Standard | Standard    | 5     | 99.97 $\pm$ 0.04 | 99.97 $\pm$ 0.04 |
| DenseNet121      | FULL           | Standard | Standard    | 5     | 99.96 $\pm$ 0.04 | 99.96 $\pm$ 0.04 |
| VGG16            | RANDBG         | Standard | Standard    | 5     | 99.96 $\pm$ 0.03 | 99.96 $\pm$ 0.03 |
| DenseNet121      | INDOOR         | Standard | Standard    | 5     | 99.96 $\pm$ 0.03 | 99.96 $\pm$ 0.03 |
| VGG16            | INDOOR         | Standard | Standard    | 5     | 99.96 $\pm$ 0.02 | 99.96 $\pm$ 0.02 |
| ResNet50         | RANDBG         | Standard | Progressive | 5     | 99.96 $\pm$ 0.02 | 99.96 $\pm$ 0.02 |
| VGG16            | FULL           | Standard | Standard    | 5     | 99.93 $\pm$ 0.07 | 99.93 $\pm$ 0.07 |
| ResNet50         | FULL           | Standard | Standard    | 5     | 99.93 $\pm$ 0.06 | 99.93 $\pm$ 0.06 |
| ResNet50         | INDOOR         | Standard | Standard    | 5     | 99.93 $\pm$ 0.05 | 99.93 $\pm$ 0.05 |
| ResNet50         | GAN            | Standard | Standard    | 5     | 99.93 $\pm$ 0.04 | 99.93 $\pm$ 0.04 |
| ResNet50         | FULL           | Standard | Progressive | 5     | 99.92 $\pm$ 0.01 | 99.92 $\pm$ 0.01 |
| VGG16            | GAN            | Standard | Standard    | 5     | 99.91 $\pm$ 0.06 | 99.91 $\pm$ 0.06 |
| ResNet50         | GAN            | Standard | Progressive | 5     | 99.91 $\pm$ 0.02 | 99.91 $\pm$ 0.02 |
| ResNet50         | INDOOR         | Standard | Progressive | 5     | 99.91 $\pm$ 0.01 | 99.91 $\pm$ 0.01 |
| VGG16            | RANDBG         | Standard | Progressive | 5     | 99.89 $\pm$ 0.03 | 99.88 $\pm$ 0.03 |
| EfficientNetV2B0 | INDOOR         | Standard | Standard    | 5     | 99.88 $\pm$ 0.05 | 99.88 $\pm$ 0.05 |
| ResNet50         | REMBG only     | Standard | Standard    | 5     | 99.87 $\pm$ 0.04 | 99.87 $\pm$ 0.04 |
| EfficientNetV2B0 | RANDBG         | Standard | Standard    | 4     | 99.87 $\pm$ 0.03 | 99.87 $\pm$ 0.03 |
| VGG16            | FULL           | Standard | Progressive | 5     | 99.87 $\pm$ 0.03 | 99.87 $\pm$ 0.03 |
| DenseNet121      | GAN            | Standard | Standard    | 5     | 99.86 $\pm$ 0.05 | 99.86 $\pm$ 0.05 |
| DenseNet121      | REMBG only     | Standard | Standard    | 5     | 99.86 $\pm$ 0.04 | 99.86 $\pm$ 0.04 |
| VGG16            | INDOOR         | Standard | Progressive | 5     | 99.86 $\pm$ 0.03 | 99.86 $\pm$ 0.03 |
| EfficientNetV2B0 | GAN            | Standard | Standard    | 5     | 99.85 $\pm$ 0.05 | 99.85 $\pm$ 0.05 |
| VGG16            | GAN            | Standard | Progressive | 5     | 99.85 $\pm$ 0.02 | 99.85 $\pm$ 0.02 |
| ResNet50         | No shift       | Standard | Standard    | 5     | 99.85 $\pm$ 0.02 | 99.85 $\pm$ 0.02 |
| VGG16            | REMBG only     | Standard | Standard    | 5     | 99.84 $\pm$ 0.03 | 99.84 $\pm$ 0.03 |
| VGG16            | No alpha       | Standard | Standard    | 5     | 99.84 $\pm$ 0.03 | 99.84 $\pm$ 0.03 |
| EfficientNetV2B0 | FULL           | Standard | Standard    | 5     | 99.83 $\pm$ 0.05 | 99.83 $\pm$ 0.05 |
| VGG16            | Ratio 1x       | Standard | Standard    | 5     | 99.82 $\pm$ 0.05 | 99.82 $\pm$ 0.05 |
| VGG16            | No shift       | Standard | Standard    | 5     | 99.82 $\pm$ 0.04 | 99.82 $\pm$ 0.04 |
| DenseNet121      | RANDBG         | Standard | Progressive | 5     | 99.82 $\pm$ 0.04 | 99.82 $\pm$ 0.04 |
| ResNet50         | Style transfer | Standard | Standard    | 5     | 99.81 $\pm$ 0.06 | 99.81 $\pm$ 0.06 |
| ResNet50         | Ratio 1x       | Standard | Standard    | 5     | 99.81 $\pm$ 0.06 | 99.81 $\pm$ 0.06 |
| DenseNet121      | No alpha       | Standard | Standard    | 5     | 99.80 $\pm$ 0.07 | 99.80 $\pm$ 0.07 |
| VGG16            | Style transfer | Standard | Standard    | 5     | 99.80 $\pm$ 0.06 | 99.80 $\pm$ 0.06 |
| ResNet50         | No alpha       | Standard | Standard    | 5     | 99.80 $\pm$ 0.03 | 99.80 $\pm$ 0.03 |

Table A1 (continued).

| Backbone         | Ablation       | Head            | Tune        | Folds | Acc. (%)         | Macro- $F_1$ (%) |
|------------------|----------------|-----------------|-------------|-------|------------------|------------------|
| DenseNet121      | INDOOR         | Standard        | Progressive | 5     | 99.80 $\pm$ 0.01 | 99.80 $\pm$ 0.01 |
| DenseNet121      | FULL           | Standard        | Progressive | 5     | 99.78 $\pm$ 0.02 | 99.78 $\pm$ 0.02 |
| DenseNet121      | Ratio 1x       | Standard        | Standard    | 5     | 99.77 $\pm$ 0.07 | 99.77 $\pm$ 0.07 |
| DenseNet121      | GAN            | Standard        | Progressive | 5     | 99.76 $\pm$ 0.02 | 99.76 $\pm$ 0.02 |
| EfficientNetV2B0 | FULL           | Attention       | Standard    | 5     | 99.74 $\pm$ 0.03 | 99.74 $\pm$ 0.03 |
| DenseNet121      | No shift       | Standard        | Standard    | 5     | 99.73 $\pm$ 0.06 | 99.73 $\pm$ 0.06 |
| DenseNet121      | Style transfer | Standard        | Standard    | 5     | 99.66 $\pm$ 0.08 | 99.66 $\pm$ 0.08 |
| MobileNetV3Small | RANDBG         | Standard        | Standard    | 5     | 98.68 $\pm$ 0.03 | 98.68 $\pm$ 0.03 |
| MobileNetV3Small | INDOOR         | Standard        | Standard    | 5     | 98.66 $\pm$ 0.14 | 98.66 $\pm$ 0.14 |
| MobileNetV3Small | GAN            | Standard        | Standard    | 5     | 98.36 $\pm$ 0.05 | 98.36 $\pm$ 0.05 |
| MobileNetV3Small | FULL           | Standard        | Standard    | 5     | 98.35 $\pm$ 0.05 | 98.35 $\pm$ 0.05 |
| EfficientNetV2B0 | Ratio 1x       | Standard        | Standard    | 5     | 98.26 $\pm$ 3.30 | 98.26 $\pm$ 3.31 |
| EfficientNetV2B0 | No alpha       | Standard        | Standard    | 5     | 98.25 $\pm$ 3.32 | 98.24 $\pm$ 3.35 |
| EfficientNetV2B0 | FULL           | Spatial pooling | Alt.        | 5     | 96.15 $\pm$ 2.05 | 96.14 $\pm$ 2.05 |
| EfficientNetV2B0 | Style transfer | Standard        | Standard    | 5     | 95.42 $\pm$ 3.93 | 95.40 $\pm$ 3.94 |
| EfficientNetV2B0 | No shift       | Standard        | Standard    | 5     | 95.20 $\pm$ 4.20 | 95.20 $\pm$ 4.20 |
| MobileNetV3Small | REMBG only     | Standard        | Standard    | 5     | 94.85 $\pm$ 7.70 | 94.82 $\pm$ 7.78 |
| EfficientNetV2B0 | REMBG only     | Standard        | Standard    | 5     | 94.72 $\pm$ 2.87 | 94.70 $\pm$ 2.88 |
| EfficientNetV2B0 | FULL           | Attention       | Progressive | 5     | 93.33 $\pm$ 0.56 | 93.32 $\pm$ 0.56 |
| EfficientNetV2B0 | FULL           | Standard        | Progressive | 5     | 93.21 $\pm$ 0.31 | 93.19 $\pm$ 0.30 |
| EfficientNetV2B0 | RANDBG         | Standard        | Progressive | 5     | 93.12 $\pm$ 0.40 | 93.11 $\pm$ 0.40 |
| EfficientNetV2B0 | INDOOR         | Standard        | Progressive | 5     | 93.05 $\pm$ 0.33 | 93.03 $\pm$ 0.32 |
| EfficientNetV2B0 | GAN            | Standard        | Progressive | 5     | 92.90 $\pm$ 0.28 | 92.87 $\pm$ 0.29 |
| MobileNetV3Small | Style transfer | Standard        | Standard    | 5     | 91.64 $\pm$ 9.04 | 91.56 $\pm$ 9.16 |
| MobileNetV3Small | No alpha       | Standard        | Standard    | 5     | 91.56 $\pm$ 9.02 | 91.51 $\pm$ 9.09 |
| MobileNetV3Small | Ratio 1x       | Standard        | Standard    | 5     | 88.48 $\pm$ 8.83 | 88.39 $\pm$ 8.92 |
| MobileNetV3Small | No shift       | Standard        | Standard    | 5     | 88.47 $\pm$ 8.90 | 88.39 $\pm$ 8.98 |
| MobileNetV3Small | GAN            | Standard        | Progressive | 5     | 82.93 $\pm$ 0.41 | 82.85 $\pm$ 0.40 |
| MobileNetV3Small | FULL           | Standard        | Progressive | 5     | 82.86 $\pm$ 0.70 | 82.74 $\pm$ 0.71 |
| MobileNetV3Small | INDOOR         | Standard        | Progressive | 5     | 82.68 $\pm$ 0.62 | 82.57 $\pm$ 0.64 |
| MobileNetV3Small | RANDBG         | Standard        | Progressive | 5     | 82.63 $\pm$ 0.78 | 82.51 $\pm$ 0.75 |
| DenseNet121      | No synth       | Standard        | Standard    | 5     | 56.70 $\pm$ 4.59 | 51.23 $\pm$ 3.03 |
| ResNet50         | No synth       | Standard        | Standard    | 5     | 52.79 $\pm$ 5.99 | 47.08 $\pm$ 5.44 |
| VGG16            | No synth       | Standard        | Standard    | 5     | 45.70 $\pm$ 5.25 | 40.11 $\pm$ 7.36 |
| EfficientNetV2B0 | No synth       | Standard        | Standard    | 5     | 43.10 $\pm$ 6.81 | 39.86 $\pm$ 6.60 |
| MobileNetV3Small | No synth       | Standard        | Standard    | 5     | 41.81 $\pm$ 4.94 | 36.74 $\pm$ 4.60 |