# Metamorphic Testing for Pose Estimation Systems

Matías Duran * Thomas Laurent * Ellen Rushe † Anthony Ventresque *

* *SFI Lero & School of Computer Science and Statistics*
*Trinity College Dublin* - Dublin, Ireland
{mduran | tlaurent | anthony.ventresque}@tcd.ie

† *School of Computing*
*Dublin City University* - Dublin, Ireland
ellen.rushe@dcu.ie

***Abstract*—Pose estimation systems are used in a variety of fields, from sports analytics to livestock care. Given their potential impact, it is paramount to systematically test their behaviour and potential for failure. This is a complex task due to the oracle problem and the high cost of manual labelling necessary to build ground truth keypoints. This problem is exacerbated by the fact that different applications require systems to focus on different subjects (e.g., human versus animal) or landmarks (e.g., only extremities versus whole body and face), which makes labelled test data rarely reusable. To combat these problems we propose MET-POSE, a metamorphic testing framework for pose estimation systems that bypasses the need for manual annotation while assessing the performance of these systems under different circumstances. MET-POSE thus allows users of pose estimation systems to assess the systems in conditions that more closely relate to their application without having to label an ad-hoc test dataset or rely only on available datasets, which may not be adapted to their application domain. While we define MET-POSE in general terms, we also present a non-exhaustive list of metamorphic rules that represent common challenges in computer vision applications, as well as a specific way to evaluate these rules. We then experimentally show the effectiveness of MET-POSE by applying it to Mediapipe Holistic, a state of the art human pose estimation system, with the `FLIC` and `PHOENIX` datasets. With these experiments, we outline numerous ways in which the outputs of MET-POSE can uncover faults in pose estimation systems at a similar or higher rate than classic testing using hand labelled data, and show that users can tailor the rule set they use to the faults and level of accuracy relevant to their application.**

***Index Terms*—Pose estimation, Metamorphic testing,**

## I. INTRODUCTION

*Pose Estimation Systems* has applications in medicine [1], sign language recognition [2] and high stakes sports events [3], requiring them to be well-tested in order to provide a substantiated assessment of the correct behaviour when applied to sensitive domains. Such systems need to perform correctly and as expected under a variety of conditions. This work aims to provide practitioners with a means of assessing this.

In this work we propose MET-POSE, a metamorphic testing framework to test pose estimation systems without the significant cost of manual data labelling. Additionally, we propose a non exhaustive set of metamorphic rules for MET-POSE, including flexible metrics to assess violations. These rules allow practitioners to apply various commonly encountered image changes and can easily be extended by users, should they want to explore different aspects of the system.

We apply MET-POSE to Mediapipe Hollistic [4], a widely used state-of-the-art pose estimation system, on datasets from the literature used to train and assess human pose estimation systems in different domains. We show that our proposed framework can find numerous faulty outputs from the system. Results show that MET-POSE provides results on par with classic, human annotated ground truth-based testing on the `FLIC` dataset. Furthermore, we illustrate how analysis of the results can highlight elements of the input that impact the system's performance. Such analysis can then help practitioners better understand the settings under which the system functions properly, and can thus be used with confidence.

The remainder of this paper is organised as follows. First, Section II provides an overview of the context and concepts that underpin this work. Next, Section III describes our proposed metamorphic testing framework for pose estimation systems, and Section IV describes a non-exhaustive set of possible metamorphic relations for this system. Section V describes the experiments performed to evaluate the framework, and Section VI describes and analyses the results of these experiments. Section VII gives an overview of related work, while Section VIII describes threats to the validity of this work, and steps taken to address them. Finally, Section IX concludes the paper and proposes avenues for future work.

## II. BACKGROUND

This section provides an overview and definition of the concepts used in this work. First, Section II-A defines human pose estimation systems, which the proposed framework tests, then Section II-B defines challenges specific to testing Machine Learning (ML) systems, and Section II-C explains the idea of metamorphic testing, the basis of the proposed framework.

### A. Pose Estimation

Pose estimation is the task of estimating the locations of different landmarks on a subject from images or video frames, with the overall aim of providing an overview of their pose. This task is important to a wide variety of fields including human activity recognition [5], sign language recognition [2], and sports analytics [3]. It is typically solved using DL-based regression algorithms which estimate the coordinates of each body part. One of the most common open-source

This work was supported in part with the financial support of grant 13/RC/2094_2 to Lero - the Research Ireland Research Centre for Software.

pose estimation frameworks available is MediaPipe's Pose Landmark Detection system [6], which is often incorporated directly into larger ML frameworks for a variety of tasks. This is typically done without fine-tuning, as this model is trained on a large and diverse set of people – often a far larger number of individuals than would be available for the lower-resource tasks on which this model is typically applied. It is thus important that these pose estimation systems be tested under different conditions when integrated into different systems that target different use cases and thus will provide images of different natures (e.g., dynamic applications will feed the pose estimation system images that are more blurry).

### B. Challenges in Testing ML-based Systems

Deep Learning (DL) research has gained enormous momentum over the last decade, with a majority of the development in this area being centred around computer vision. However, despite their success, these deep architectures have also demonstrated harmful decision-making [7]–[9], bias [10] and a lack of "understanding" of common-sense concepts such as basic spatial relations [11], [12]. These issues are made all the more challenging to identify given that DL methods lack interpretability, with the steps that lead to a particular decision often being unclear. This lack of transparency means that it is crucial that these systems are systematically tested and their results compared to their expected behaviour in order to understand their sensitivities, along with the conditions under which they can be expected to behave correctly.

Another aspect that complicates the testing of complex systems is that they often present what is known as the *Oracle Problem* [13], which describes a scenario where the correct output of the system under test is not known. Sometimes there are no automated ways to compute this oracle, or the act of querying this oracle can be prohibitively expensive. In the case of pose estimation, labelling keypoints in videos is extremely time consuming, making it expensive to collect this form of ground truth data for many applications. This complicates the testing of these systems and requires techniques that can assess the correct functioning of the system without needing an oracle for the correctness of a single execution of the system.

Though testing techniques for computer vision-based classification systems have been extensively explored [14]–[16], less attention has been given to more complex computer vision tasks such as pose estimation. For example, while MediaPipe does evaluate its pose estimation framework for bias related to the demographics of individuals – an admirable activity – this analysis is expensive as it requires ground truth information. Additionally, though their evaluation states that degradation can be expected as video quality and lighting gets worse, a comprehensive testing procedure is not outlined. For many applications, it is crucial that we understand the *specific* parameters within which a system can be expected to operate accurately, especially where mistakes are costly. There has been some work in this area [17]

### C. Metamorphic testing

*Metamorphic Testing* ($MT$) has been used to address the oracle problem for both classical programs [18] and ML applications [15] and has demonstrated promising results at a relatively low cost. $MT$ relies on *metamorphic rules* – relations between certain changes to the input of the system (e.g., doubling an integer input) and their effect on its output (e.g., the output should be doubled too) – to circumvent the oracle problem. If one of these relations is violated for a given input, then the system is not behaving correctly. Note that the violation of the metamorphic rule can be verified without knowing what the correct output for either of the inputs is, i.e., without an oracle for the correctness of each output.

This method is well suited for testing pose estimation systems as it does not require ground truth labels, which come from expensive manual annotation of data. Additionally, since the inputs can be modified in numerous ways, it can help understand the boundaries of the input space under which the system performs as expected.

## III. METAMORPHIC TESTING FOR POSE ESTIMATION (MET-POSE)

This section first formally defines the problem of testing pose estimation systems (III-A) and then introduces MET-POSE (III-B).

### A. Problem definition

MET-POSE aims to test pose estimation systems at large. Given an input image $img$, a pose estimation system returns an output $O$ which is a list of sets of keypoints: $O = [KP_0, KP_1, ..., KP_n]$ with $n \in \mathbb{N}$ or $O = []$. Each list of keypoints $KP_i$ corresponds to the keypoints of a subject detected in the image by the system, i.e., $\forall i \in \{0..n\}, KP_i = kp_0, ..., kp_p$, with each keypoint $kp_j$, identifying a landmark of subject $i$ and its coordinates.

Testing these systems is an inherently complex task, as they often lack exact requirements. For example, the list of landmarks that should be found on a person in the task of human pose estimation is not defined in a consistent manner with each system defining its own. Similarly, the way to account for occlusion in an image or to process images containing no or multiple subjects is often not defined. This lack of standard requirements contributes to the oracle problem and to the cost of testing pose estimation systems, as each system requires its own hand labelled data for evaluation.

Additionally, current systems that rely on pose estimation are mostly built using deep neural networks-based pose estimation system. These models are typically trained on large, diverse datasets to facilitate generalisation. Given the "generalised" nature of these models and high cost of labelling pose data, these models are typically used "out-of-the-box", without fine-tuning them for the particular application of the system they are used in, and even sometimes without domain-specific testing. Additionally, different use cases and environmental factors can challenge a pose estimation system in different ways, for example: low lighting, different camera angles, or

motion blur. We consider that the problem of testing pose estimation systems should be embedded in a particular use case and consider these application-specific conditions.

MET-POSE aims at tackling these problems with the testing of pose estimation systems. It aims at providing a general, flexible, and tunable testing method for pose estimation systems that does not require hand labelled ground truth test data. MET-POSE thus offers a testing method that lets users tests these systems for their particular use cases.

### B. Approach overview

We aim for MET-POSE to be a general framework so we present it here in broad terms. A specific example of how the framework can be applied is detailed in Section IV. MET-POSE is a metamorphic testing framework for pose estimation systems. It thus relies on *metamorphic rules* based on images and keypoints. Each rule $\mathcal{M}$ is defined by two elements:

- A *transformation* $\mathcal{M}.trans$, that defines a modification to apply to an image, e.g., making the image greyscale. Given an original test image $img_{orig}$, $\mathcal{M}.trans$ produces a modified test image $img_{mod} = \mathcal{M}.trans(img_{orig})$.
- A *relation* $\mathcal{M}.rel$, that defines a property between the System Under Test (SUT)'s output keypoints for $img_{orig}$ and its output keypoints for $img_{mod}$ that should appear if the system functions correctly. The simplest relation is the identity relation, i.e., the system should output the same keypoints for $img_{orig}$ and $img_{mod}$. If this relation is not followed (i.e., the property does not appear), then the rule is said to be violated.

A violation of a metamorphic rule on an image indicates that the pose estimation system is providing incorrect output on $img_{orig}$, on $img_{mod}$, or on both. The severity of the system's error can vary, i.e., the output can be more incorrect in some cases than others. This is something that MET-POSE takes into account by returning the severity of a violation, which can then be considered by, for example, using different error thresholds to decide if a violation constitutes a test failure.

Fig. 1 gives an overview of the MET-POSE framework. MET-POSE takes as input the SUT, a set of metamorphic rules, and a set $IMG_{orig}$ of test images. MET-POSE applies the transformation of each rule to each test image in $IMG_{orig}$, and for each pair of original and modified images it checks whether the rule is violated, and to what degree. For each input image $img_{orig} \in IMG_{orig}$, and each metamorphic rule $\mathcal{M}$ in the input set, the output of MET-POSE is: • a modified image $\mathcal{M}.trans(img_{orig})$, • whether the pair of images violates $\mathcal{M}.rel$, • how severely the pair violates $\mathcal{M}.rel$ if it does. Users are then free to analyse this output based on the quality constraints of their application. Some use cases, for example, might not require that small violations be considered failures if high precision is not required.

MET-POSE is agnostic to the way rules are defined and how the relations are assessed. The next section details examples of rules and a possible error metric-based technique to assess their relations.

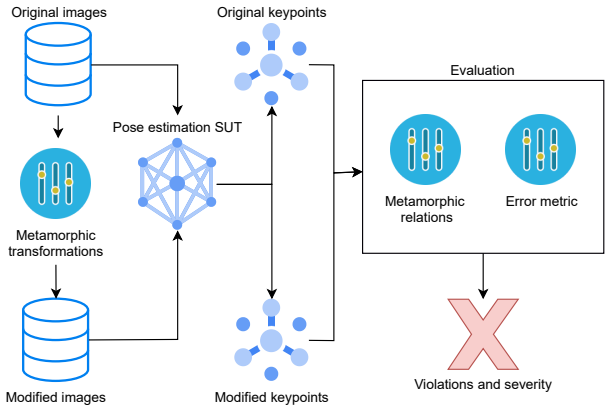



Fig. 1: Overview diagram of MET-POSE

## IV. PROPOSED METAMORPHIC RULES

MET-POSE can be used with any metamorphic rule that defines a transformation $trans$ of an input image $img$ and a relation $rel$ between the keypoints of $img$ and $trans(img)$. This section describes a non-exhaustive set of rules that we propose to evaluate MET-POSE with. Section IV-A describes the transformations and relations of these rules and Section IV-B defines a possible error-metric based mechanism to assess violations of the rules. Section V-D defines the exact metric used in this work.

### A. Transformations and Relations

This section describes a varied set of rules for MET-POSE that cover a range of transformations that are often encountered in the applications of pose estimation systems. In each different application of pose estimation, it can be useful to compare the effect of various modifications to the input images to infer which specific modifications lead to higher output inaccuracies. This is particularly important when there is a possibility of collecting new data for a given task, as these data quality issues can be potentially mitigated or minimised.

Fig. 2 shows the effect of some of the transformations used in this work. We detail here the proposed rules and settings used in the experiment. The rules can be classified based on their transformations:

- Spatial transformations.
  These rules focus on the shape or orientation of the image:
  - **Identity** (`Id`): does not modify the image. $rel$: the keypoints on both images should be equal. Checks that the SUT is deterministic.
  - **Stretch** ($\texttt{Stch}_{h,w}$): stretch the image both on the vertical (by a factor of $h$) and horizontal ($w$) axes. $rel$: the keypoints should keep at the same relative coordinates (0,0 is at bottom left of the image, and 1,1 at top right) in the stretched and original images.
  - **Mirror** (`Mirr`): mirrors the image along the horizontal ($\texttt{Mirr}_h$), vertical ($\texttt{Mirr}^v$) or both ($\texttt{Mirr}^v_h$) axes. $rel$:

keypoints should be mirrored in the chosen axes. Mimics situations such as video-calls where the image can be mirrored.

  - **Rotation** ($\texttt{Rot}^{c}_{\omega}$): rotates the image $\omega$ degrees around a given centre point $c$. $rel$: keypoints on the modified image should be equal to the keypoints on the original image rotated $\omega$ degrees around $c$.

- Image quality transformations.
  These rules focus on problems often encountered with the quality of images provided as input to pose estimation systems and challenges often encountered in practice. Fig. 3 provides an example of such problems (motion blur and pixelation) for sports analytics.
  The relation for these rules is that the keypoints should be unchanged after the transformation.
  - **Resolution** ($\texttt{Res}_{factor}$): multiplies the resolution of the image by a factor $0 <= factor < 1$. Lower resolutions of images could be a concern for applications using older or cheaper image capturing devices.
  - **Gamma correction** ($\texttt{Gamma}_{\gamma}$): applies gamma correction [19] to the image with a gamma $\gamma$. Brightness can present high variations in real world applications.
  - **Brightness Scaling** ($\texttt{Bright}^{m}_{a}$): for each pixel and each channel value $v^{c}_{i,j}$ returns $a + m * v^{c}_{i,j}$.
  - **Bilateral Filtering** ($\texttt{Bilat}^{size}_{str}$): applies a bilateral filter [20] of strength ($str$) and size ($size$) to the image, preserving edges of the image but reducing textures. By preserving the edges of objects but reducing textures, this rule checks if the SUT is too reliant on textures instead of the shape of objects, as shown by Geirhos et al. [21].
  - **Motion** ($\texttt{Motion}^{dir}_{k\_s}$): applies motion blur to the image with a given kernel size and direction. Motion blur is often seen on images used for pose estimation, as Fig. 3 shows. This is especially true when pose estimation is used for hands [22].

- Colour-space transformations.
  These rules change the colours of the image in a way that does not affect the clarity of the subjects in order to ensure that the pose estimation system is properly recognising their pose, and not relying on features unrelated to the pose, such as colours.
  The relation for these rules is that the keypoints should be unchanged after the transformation. All colour transformations test the SUT's over-reliance on colour, which can affect performance when the system is used in a new context [23].
  - **Greyscale** ($\texttt{Grey}$): turns the image into a greyscale (only shades of grey) image.
  - **Colour wheel** ($\texttt{CWheel}_{\theta}$): changes colours in the image by applying a rotation $\theta$ on the hue colour wheel value of a HSV [24] encoding of the input image.
  - **Colour channels** ($\texttt{Cchans}^{encoding}_{[factor1,factor2,factor3]}$): multiplies each of the 3 colour channels in the $encoding$ (RGB, BGR or XYZ) of the image by a given factor.

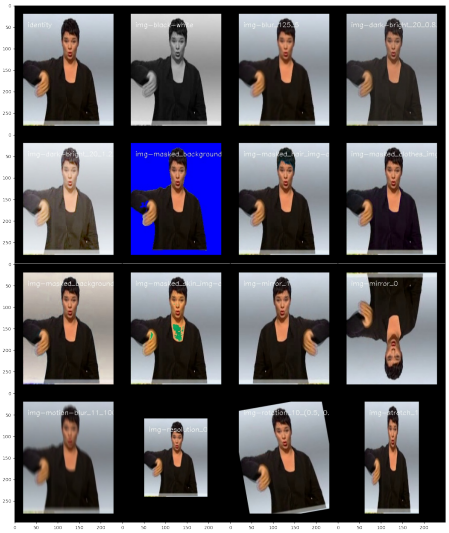


Fig. 2: Example image modified by various rules

Fig. 3: Example inputs from a sports analytics application, showing motion blur and rotation; occlusion; and pixelation

To understand the influence of different zones of the image, we introduce "filtered" versions of certain rules: $\texttt{Flt}(\texttt{rule}, \texttt{zone})$. The filtered version of a rule only applies the transformation to a particular zone of the image, e.g., the background of the image (all non-subject zones).

The last rule we propose in this work is the **Colour fill** ($\texttt{Cfill}_{colour}$). This rule is only used in a filtered way, and therefore fills only a segment of the image a certain colour. Its relation is that keypoints should remain unchanged.

### *B. Error Metric-based Relation Assessment*

This section introduces an error metric-based technique to assess violations of rules used in MET-POSE. All the rules proposed in this work either expect no change in keypoints or modify the position of the expected output keypoints between the original and modified. However, different rules, e.g., those involving removing or adding landmarks to the image, could

also modify the keypoints that the system is expected to detect. Assessing the degree of violation of these different types of rules would require different approaches, which could all be implemented as an error metric (e.g., counting the number of added landmarks that were not detected by the SUT).

For each rule, the user can provide a metric $\texttt{Err}(O_{expected}, O_{mod})$, where $O_{expected}$ is the expected position of the keypoints on the modified image according to the metamorphic rule's relation. This metric should be tailored to each relation and to the types of errors considered (e.g., focusing only on hand keypoints). MET-POSE lets the user define an error threshold $t_{err}$ over which the error is not acceptable. A test then passes iff:

$$\texttt{Err}(O_{expected}, O_{mod}) < t_{err}$$

Both the error metric and error threshold are definable by the user, giving them full control over the aspects of the system's output they want to test (e.g., ignoring particular keypoints), along with the sensitivity of those tests. Section V-D details the human pose estimation-oriented metric used in the experiments, and a comparison of results for different error thresholds.

## V. EXPERIMENTS

This section details the experiments conducted to illustrate how MET-POSE can be used and evaluate how well it can find faults in a pose estimation system without using ground truth keypoints. Section V-A details the research questions we explore with these experiments. Sections V-B and V-C detail the pose estimation system under test and datasets used in the experiments, and Section V-D defines the error metric used to evaluate rule violations.

### A. Research questions

The experiments in this work aim to answer the following research questions (RQs):

- **RQ1**: Does MET-POSE find faults?
  This RQ considers whether MET-POSE can find faults in pose estimation systems. In order to explore this question we consider the number of rule violations found by MET-POSE using different subsets of the rules described in Section IV-A and different error thresholds.
- **RQ2**: How different are the results of MET-POSE and classic, ground truth-based testing?
  This RQ focuses on whether the faults found by MET-POSE correspond to those found using classic, ground truth-based testing using the same error metric and error threshold as MET-POSE. This research question considers the same rule sets and error thresholds as RQ1 and is composed of three sub-RQs:
  - **RQ2.1**: Do both testing methods lead to a similar number of failures?
    This RQ first assesses whether both testing methods uncover the same number of failing test cases, a test case being considered as an input image for both methods.
  - **RQ2.2**: Do the same test images make the system fail with both methods?
    This RQ explores whether the particular images that lead to a failure are the same for both methods, i.e., if they expose the same system failures.
  - **RQ2.3**: How large are the errors found by both methods?
    Finding larger (w.r.t. the chosen error metric) errors in addition to small errors would show that a testing method finds faults in the SUT with a stronger effect on the output.
- **RQ3**: What do the different proposed metamorphic rules bring to the method?
  This RQ explores the contribution of the different rules proposed in Section IV-A[1] to MET-POSE. It is composed of two sub-RQs:
  - **RQ3.1** Are there subsumption relationships amongst the proposed rules?
    This RQ explores possible subsumption relations between the different proposed rules. A rule subsumes another rule if it finds the failure-inducing images another rule does, making the second rule redundant.
  - **RQ3.2** Do the different metamorphic relations reveal the same types of faults?
    This RQ assesses whether the different metamorphic rules find the same failure-inducing images, which would indicate the violations of these rules could have their roots in the same fault in the system.

### B. System Under Test

In order to explore the three RQs defined in Section V-A, we applied MET-POSE to *Mediapipe Holistic* [4], a state of the art human pose estimation system. Its underlying DL model is BlazePose GHUM 3D [25]. The BlazePose [26] model is a lightweight convolutional neural network that predicts the co-ordinates of body parts.

*Holistic* is composed of three sub-models, each one responsible for the keypoints of the: face (468 landmarks), hands (21 landmarks each), and overall body pose (33 landmarks).

We used *Holistic* in static image mode. Using the system in this mode provides a deterministic output, as verified by the use of the `Id` rule (i.e. there is no difference between two different generations without a transform). This ensures that no noise from non-determinism is integrated into the error metric.

### C. Datasets

We tested the SUT with MET-POSE with two datasets:

- `PHOENIX` [27] is a dataset created for sign language recognition. It contains $947,756$ video frames from a signed German weather forecast. Given that this dataset is used in the field of sign-language recognition and translation, the dataset contains manually labelled annotations in

[1]Note that these rules are general examples that cover general (human) pose estimation challenges and that, as described in Section III-B, MET-POSE can be used with any metamorphic rules, allowing users to tailor the rules they use to the particular application they want to test pose estimation for.

the form of written text translation of signs, but no ground truth keypoints. The images are in a very controlled environment, with all video frames having a similar plain background, with subjects wearing plain black clothes, and without large changes in the subject's distance from the camera or in the angles of their poses.

- FLIC [28] is a dataset containing $4,552$ movie frames curated from popular movies. All images include one or multiple subjects (people) as it is specifically a human pose estimation benchmark. The environment of this dataset is less controlled, with situations changing drastically depending on the movie scene images were taken from. This provides a contrast to the controlled environment of PHOENIX. FLIC contains ground truth ($gt$) manual annotations of 11 keypoints obtained from multiple annotators through Amazon Mechanical Turk. We have mapped a subset of keypoints from *Mediapipe* to the corresponding keypoints of the FLIC $gt$.

Both of these datasets use publicly available images. FLIC was specifically curated to benchmark pose estimation systems, while Phoenix requires that the person signing remains in frame, meaning that both datasets have an easily identifiable person in the frame. This is in contrast with other datasets that are built for general object classification such as COCO [29], which contain comparatively fewer human subjects or have these subjects positioned very far away from the camera. We ran our experiments on representative subsets of these two datasets in order to reduce the cost of our experiments. Specifically, for PHOENIX we used the *dev* subset ($55,775$ images). For FLIC, we used the images labelled as test, excluding the *FLIC-full* extension, i.e., $835$ images.

### D. Metamorphic Rules and Error Metrics

In order to explore the research questions that Section V-A details, we applied MET-POSE using a set of the rules that Section IV-A describes using a range of configurations of these relations. Table I details this set of configurations, which is denoted by $AllRels$.

Not all of the relations and configurations in $AllRels$ are relevant to the data in PHOENIX and FLIC and to testing *Holistic* without finetuning, leading to artificially high error rates. We thus also use a restricted set of relations and settings, highlighted in bold in Table I and denoted by $SubRels$. Additionally, we consider Grey and $\texttt{Mirr}_h$ in order to assess whether MET-POSE can find problems with *concept understanding* in these systems – an area previously shown to be faulty in other deep learning-based models [11]. Finally, we determine whether MET-POSE can surface particular faults in the pose estimation system such as over-reliance on colour (which can vary substantially based on the lighting and other conditions) or orientation (e.g., dealing with left and right handed signers in PHOENIX, or with mirrored input from online calls)– two features that should not impact its output.

To assess violations of all rules used in the experiments we use the same error metric $Err_{lmks}$ defined as:

$$Err_{lmks}(O_{expct}, O_{trans}) = \begin{cases} \text{if } O_{expct} = \emptyset \veebar O_{trans} = \emptyset, \inf \\ \text{elif } O_{expct} = \emptyset \wedge O_{trans} = \emptyset, 0 \\ \text{else, } \underset{lmk \in lmks}{\text{Med}} L2_{MP}(kp_{expct}, kp_{trans}) \end{cases}$$

If the pose estimation system only returns keypoints on the original image or the modified image, $Err_{lmks}$ returns an infinite value. Indeed, the transformations used in these experiments do not modify which landmarks the system should detect on the image, at most they change the expected position of the returned keypoints. Failing to detect any landmarks on just one of the pair of images is thus the worst violation of a rule possible. However, if the pose estimation system returns no keypoints for both images it displays a coherent behaviour w.r.t. the metamorphic rule and the rule is not violated. Note that, in general, not returning keypoints does not denote a fault in itself as an image could contain no visible person. However, FLIC and PHOENIX both only contain images containing people so $O_{orig} = \emptyset$ does denote a fault, albeit not one that would violate our proposed rules.

If the pose estimation system returned keypoints for both images then $Err_{lmks}$ returns the median of the distance between the expected keypoint (following the metamorphic rule's relation and the keypoints returned by the system on the original image) and the keypoint returned by the system on the modified image for each landmark $lmk$. Using the median as a measure of the overall error of the system on an image attenuates the effect of outliers on the error value. Indeed, if only a single keypoint's coordinates are extremely inaccurate, many applications would still work. Note however that the error of all keypoints in an image could be aggregated in other ways (min, max, ...) depending on the needs of the use case the pose estimation system is tested for. Here the distance used, $L2_{MP}$ is the normalised Euclidean distance as described in *Mediapipe*'s model cards [30]–[32], i.e., depending on which of the 3 groups described in section V-B the landmark belongs to, the distance is normalised by dividing by either the distance between the subject's: shoulders; irises; or their wrist and the first joint of their middle finger.

As we test *Mediapipe* without a particular use case, setting a meaningful value for the error threshold is not possible. We thus assess violations using a large range of values for the threshold and report the performance of MET-POSE using these different values.

## VI. RESULTS

This section analyses and discusses some of the results from our experiments following the research questions defined in Section V-A. The code used to generate the reported results is available in the companion repository [33].

### A. RQ1

Figure 4 shows the proportion of images in each dataset leading to a rule violation when testing *Holistic* with MET-POSE for different sets of rules and different error threshold

TABLE I: Full list of metamorphic rules settings, settings in bold are included in *SubRels*

| **Transformation** | $\in$ *SubRels* | **Configurations** |
|---|---|---|
| `Id` | ✓ | |
| $\texttt{Stch}_{1,2}$ | ✓ | (0.6, 1), (0.8, 1), (0.9, 1.1), (0.95, 1.05), (1, 1.4), (1, 1.25), **(1, 0.8)**, **(1, 0.6)**, (1.05, 0.95), (1.1, 0.9), **(1.25, 1)**, (1.4, 1) |
| $\texttt{Mirr}_h^v$ | ✓ | **horizontal**, vertical, both |
| $\texttt{Rot}_1^2$ | ✓ | **(5, (0.5, 0.5))**, **(10, (0.5, 0.5))**, (15, (0.5, 0.5)), (25, (0.5, 0.5)) |
| $\texttt{Res}_1$ | ✓ | 0.1, **0.2**, 0.3, 0.4, 0.5, 0.6, **0.7**, 0.8, 0.9, 0.95, 0.98 |
| $\texttt{Gamma}_1$ | ✓ | 0.25, **0.5**, 0.85, 0.95, 1.05, 1.15, 1.5, 1.75 |
| $\texttt{Bright}_1^2$ | ✓ | (-20, 0.8), (-20, 1.6), (0, 1.05), (0, 1.15), (20, 0.4), **(20, 0.8)**, (20, 1.2), (20, 1.6), (30, 1.15) |
| $\texttt{Bilat}_1^2$ | ✓ | **(10, 3)**, (10, 5), (10, 7), (10, 9), (30, 3), (30, 5), (30, 7), (30, 9), (50, 3), (50, 5), (50, 7), (50, 9), (80, 3), (80, 5), **(80, 7)**, (80, 9), (125, 3), **(125, 5)**, (125, 7), (125, 9), (150, 3), (150, 5), (150, 7), (150, 9), (180, 3), (180, 5), (180, 7), (180, 9) |
| $\texttt{Motion}_1^2$ | ✓ | (5, 0), (5, 40), (5, 70), (5, 100), (7, 0), (7, 40), (7, 70), (7, 100), (9, 0), (9, 40), (9, 70), (9, 100), **(11, 0)**, (11, 70), **(11, 100)** |
| `Grey` | ✓ | |
| $\texttt{Flt}(\texttt{CWheel}_1, 2)$ | ✓ | (10, skin), (30, skin), (90, skin), (-45, skin), (10, clothes), (30, clothes), (90, clothes), (-45, clothes), **(90, hair)**, **(90, background)** |
| $\texttt{Flt}(\texttt{Cchans}_1^2, 3)$ | | ([0.9, 1.1, 1.1], RGB, skin), ([1.1, 1.1, 0.9], RGB, skin), ([0.8, 1.3, 1.3], RGB, skin), ([1.3, 1.3, 0.8], RGB, skin), ([0.6, 1.4, 1], RGB, skin), ([1.4, 1, 0.6], RGB, skin), ([0.45, 1, 1.2], RGB, skin), ([1.2, 1, 0.45], RGB, skin), ([1, 1, 1], BGR, skin), ([1, 1, 1], XYZ, skin) |
| $\texttt{Flt}(\texttt{Cfill}_1, 2)$ | | ([0, 0, 255], background), ([255, 180, 120][a], background), ([33, 28, 27][b], background), ([0, 0, 255], skin), ([255, 180, 120], skin), ([33, 28, 27], skin), ([0, 0, 255], clothes), ([255, 180, 120], clothes), ([33, 28, 27], clothes) |

[a] close to skin colour on phoenix dataset [b] close to clothes colour in phoenix dataset

values when considering the body pose landmarks. These results confirm that MET-POSE can find faults in the system as it finds violations of the rules. They also confirm the central role of the rules and the error threshold value used in MET-POSE, which is why we have chosen to keep these characteristics customisable so they can be defined by users of the framework to suit their particular application.

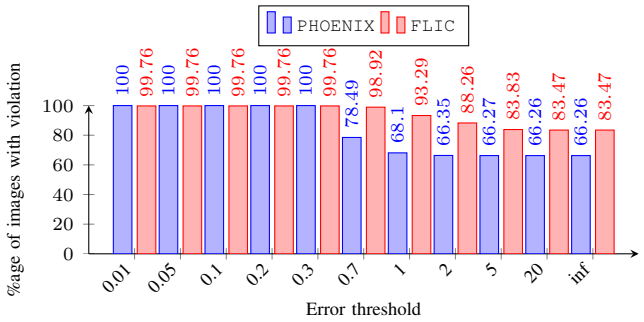
When the threshold is set very low, e.g., to 0.01[2] as the leftmost columns of the graphs illustrate, MET-POSE is very sensitive and most images lead to violations. With high error threshold values MET-POSE is much more lenient and the proportion of images leading to a violation quickly drops. However, MET-POSE still finds images that lead to an infinite error, i.e., where the pose estimation system detects the subject only in one of the original and modified images, as defined in Section V-D. These cases are always considered violations of our proposed metamorphic rules.

Figure 5 shows the proportion of images in PHOENIX leading to a rule violation with `Grey` and $\texttt{Mirr}_h$ for different error threshold values when considering the hand landmarks. These results, when contrasted with those in Figure 4, show once more the importance of adaptability in MET-POSE. While focusing on the body pose landmarks quickly exposes very few errors on FLIC when using only the `Grey` and $\texttt{Mirr}_h$, using hand landmarks exposes many more rule violations. This is consistent with the goal of PHOENIX, i.e., the fact that it is a sign language dataset, where the hands, arms, and face, are much more the focus than the general body pose.

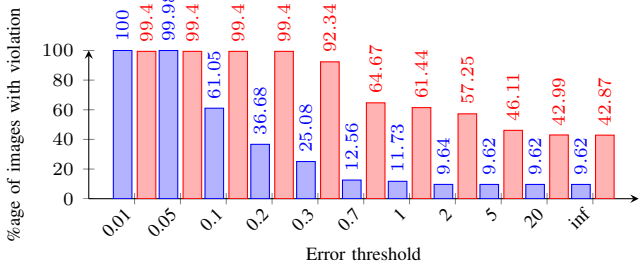
## B. RQ2

This RQ focuses on the difference between classic, ground truth based testing and MET-POSE. Thus, it relies on FLIC,

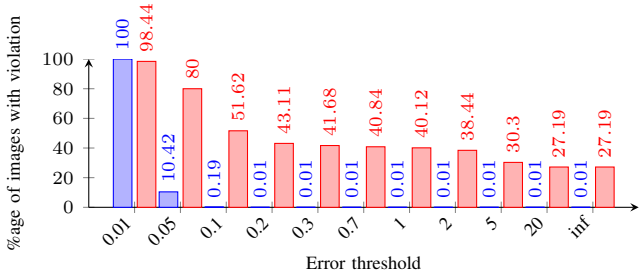
[2]In this setting, a median error of more than 1% of the distance between the subject's shoulders leads to a violation.




(a) using *AllRels*

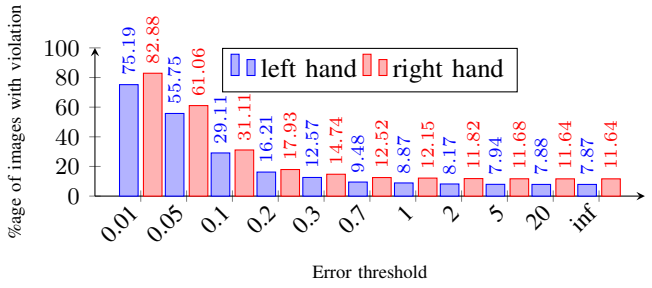



(b) using *SubRels*

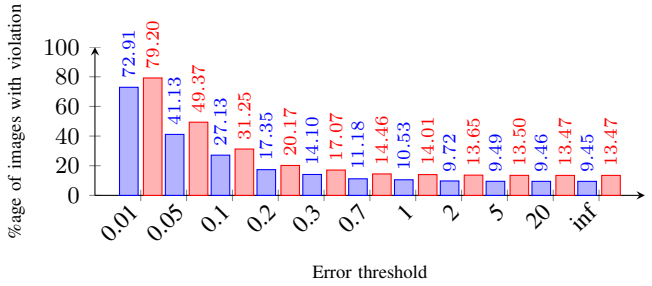



(c) using only `Grey` and $\texttt{Mirr}_h$

Fig. 4: Percentage of images leading to a rule violation with varying error thresholds using body landmarks




(a) Using only $\texttt{Mirr}_h$




(b) Using only `Grey`

Fig. 5: Percentage of images in `PHOENIX` leading to a rule violation for varying error thresholds using hand landmarks

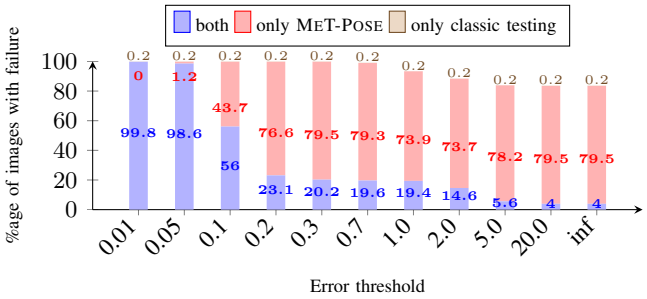
the only dataset in our experiments with ground truth labels. Figure 6 shows: the proportion of images leading to both a classic test failure and a rule violation, a classic test failure only, and a rule violation only for different error thresholds. A classic test failure is defined as a normalised distance between the ground truth keypoints and the keypoints output by the system on the original image greater than the error threshold. These results show that for low values of the error threshold, both methods perform similarly in the number of failures they find, even when using a single rule in MET-POSE. This is explained by both methods being too sensitive with that configuration and considering nearly all outputs as faults. With higher values of the error threshold we see that MET-POSE finds many more failures than classic testing when using *AllRels* or *SubRels*. The answer to RQ2.1 is thus that, overall, MET-POSE finds more failures than classic testing.

Regarding RQ2.2, we see that in all cases, there is some overlap between the images that lead to failures for each method, however each method also finds failures that the other method did not detect, with MET-POSE finding more failures when using more complete sets of rules.

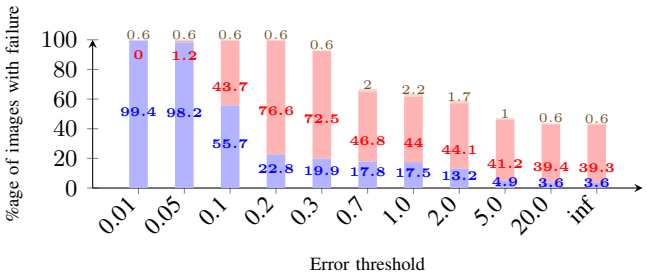
As discussed in RQ2.1, both testing methods perform similarly for smaller values of the error threshold while MET-POSE finds more failures for higher values. This indicates that MET-POSE finds larger failures overall.

### C. RQ3

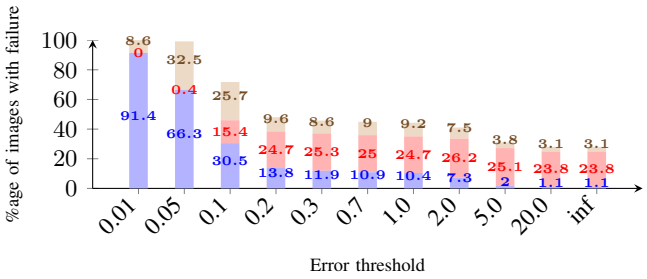
To analyse the contribution of different metamorphic relations when using MET-POSE we can look at their *subsumption rates*. A rule $\mathcal{M}_1$ subsumes a rule $\mathcal{M}_2$ if, when $\mathcal{M}_2$ detects a failure for an image then $\mathcal{M}_1$ also detected that failure.



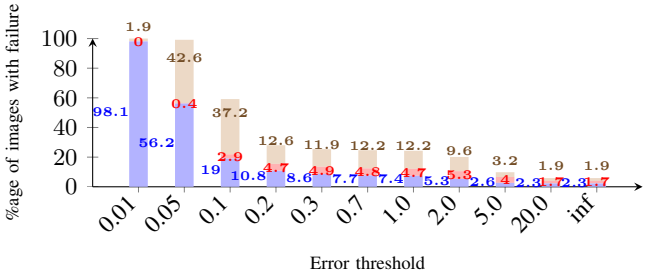

(a) with MET-POSE using *AllRels*




(b) with MET-POSE using *SubRels*




(c) with MET-POSE using `Grey`



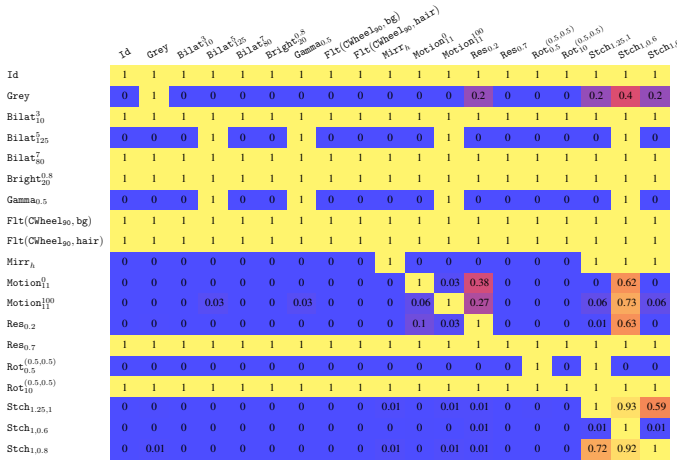

(d) with MET-POSE using $\texttt{Mirr}_h$

Fig. 6: Percentage of images in `FLIC` leading to a classic test failure or a rule violation with varying error thresholds using body landmarks

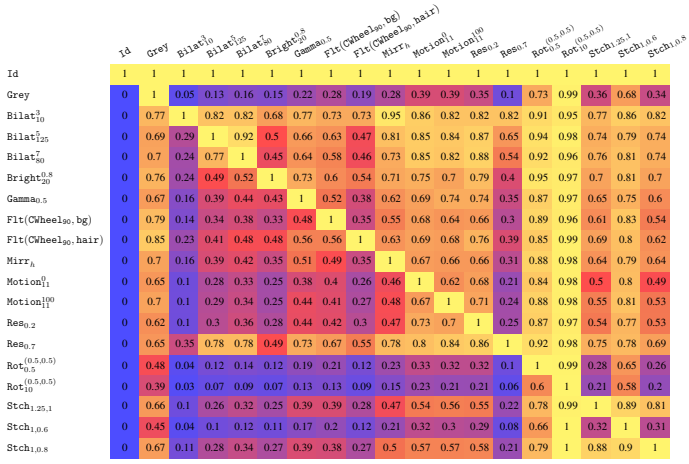
This means that running $\mathcal{M}_1$ is sufficient to detect the failures $\mathcal{M}_2$ detects. Given two rules, $\mathcal{M}_1$ and $\mathcal{M}_2$, we explore the degree to which $\mathcal{M}_1$ subsumes $\mathcal{M}_2$ using the subsumption rate $\texttt{SubRate}_{\mathcal{M}_1,\mathcal{M}_2}$:

$$\texttt{SubRate}_{\mathcal{M}_1,\mathcal{M}_2} = \begin{cases} \text{if } \mathcal{M}_1 \text{ is never violated,} 1 \\ \text{else, } \frac{\text{\# images violating } \mathcal{M}_1 \text{ and } \mathcal{M}_2}{\text{\# images violating } \mathcal{M}_1} \end{cases}$$

Figure 7 shows the subsumption rates of all pairs of rules in *SubRels* for each dataset for a set value of the error threshold



(a) PHOENIX



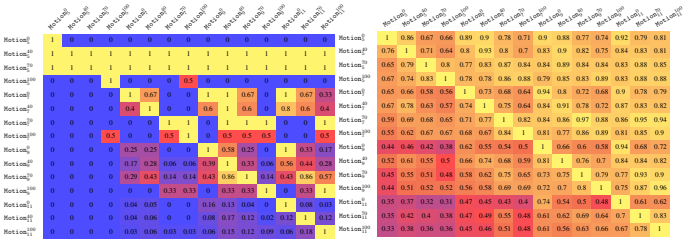
(b) FLIC

Fig. 7: $\texttt{SubRate}_{\mathcal{M}_1,\mathcal{M}_2}$, with $\mathcal{M}_1$ left and $\mathcal{M}_2$ top, for each dataset, using body pose landmarks and a 0.2 error threshold.

(0.2). These results show that the subsumption rates between the rules are very dependent on the dataset used, indicating that the different rules make different contributions to MET-POSE, i.e., they find different types of errors.

Figure 8 focuses on the different settings of the `Motion` rule and how they subsume each other. While one might expect that settings that introduce more discrete blur would clearly subsume settings that introduce stronger blur, the results do not show such a clear result. This is probably due to the complex nature of *Mediapipe Holistic*, which means that the direct effects of a transformation on the system's output are not easily predictable – further illustrating the value of the proposed testing procedure.

The data in Table II confirms that the different rules reveal different types of faults. Indeed, when using a reasonable error threshold (not so low that nearly all images lead to rule violations), most images lead to a violation of only a small number of rules. This again suggests that the different rules assess different aspects of the system as, in most cases, the same input leads to a violation only for specific rules.



(a) PHOENIX (b) FLIC

Fig. 8: Subsumption rates of various motion blur settings

## VII. Related Work

The use of machine learning systems is becoming increasingly common in many diverse fields. Testing strategies for these systems are therefore being more widely recognised as a necessity, especially in safety critical systems [34]. In particular, these systems are tested with different focuses in mind such as: robustness to degradations [17] [35] [36] [37], testing for faults when using ML frameworks [38], testing the adequacy of the datasets used to train and test these systems [39] and testing the biases embedded in these systems [8], [40].

In particular, metamorphic testing is a convenient method to circumvent the oracle problem found when attempting to test many of these systems. This method has been used to test a wide range of ML systems, including: autonomous delivery robots [41], image classification systems [42], recommender systems [43], search engines [44] and chat bots [45].

More particularly, when looking at pose estimation systems, we can find previous work focused on the pose estimation of hands [17], and on predicting the trajectory of humans, done by autonomous systems [46].

When looking at the particular model tested in this work, it has previously been explicitly evaluated on individuals from different regions of the world which has been reported in the model's corresponding Model Card [30], [47].

One of the main differences between the related work and the framework presented here is that we avoid the need for ground truth labelled data, meaning that MET-POSE can work with datasets without ground truth, which has not been done in previous metamorphic testing applied to pose estimation. We also do not focus only on robustness to degradations. We present a framework that can be used for various specific interests of the applications, which can be image degradation, but can encompass a wide range of potential concerns from fairness issues to sensitivity to image quality. Even if we do not focus on suggesting novel metamorphic rules, when novel metamorphic rules are needed for a specific application, MET-POSE can be applied to new datasets without the expense of labelling ground truth keypoints.

TABLE II: Images that violate different numbers of rules at given error thresholds in PHOENIX (PH) and FLIC (FL)

| # failed rules \ Error thresh. | 0.01 | | 0.05 | | 0.1 | | 0.2 | | 0.3 | | 0.7 | | 1.0 | | 2.0 | | 5.0 | | 20.0 | | inf | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | FL | PH | FL | PH | FL | PH | FL | PH | FL | PH | FL | PH | FL | PH | FL | PH | FL | PH | FL | PH | FL | PH |
| 0 | 2 | 0 | 2 | 0 | 2 | 0 | 2 | 0 | 2 | 0 | 9 | 11996 | 56 | 17794 | 98 | 18767 | 135 | 18813 | 138 | 18821 | 138 | 18821 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 35 | 23192 | 51 | 28258 | 77 | 28789 | 113 | 28759 | 121 | 28752 | 122 | 28752 |
| 2 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 10363 | 0 | 14486 | 52 | 11742 | 54 | 7144 | 44 | 6180 | 74 | 6171 | 84 | 6171 | 84 | 6171 |
| 3 | 1 | 0 | 1 | 0 | 1 | 1401 | 1 | 18521 | 1 | 24116 | 48 | 5985 | 52 | 1616 | 47 | 1334 | 44 | 1330 | 47 | 1329 | 49 | 1329 |
| 4 | 0 | 0 | 0 | 0 | 0 | 5129 | 0 | 15246 | 15 | 11806 | 48 | 1956 | 36 | 617 | 33 | 446 | 41 | 443 | 46 | 443 | 45 | 443 |
| 5 to 30 (25%) | 9 | 0 | 198 | 53427 | 534 | 49229 | 635 | 11644 | 643 | 5367 | 494 | 904 | 446 | 346 | 414 | 259 | 354 | 259 | 355 | 259 | 353 | 259 |
| 31 to 61 (50%) | 16 | 0 | 373 | 2668 | 206 | 16 | 136 | 1 | 118 | 0 | 111 | 0 | 104 | 0 | 92 | 0 | 59 | 0 | 36 | 0 | 36 | 0 |
| 62 to 91 (75%) | 5 | 2635 | 163 | 38 | 58 | 0 | 46 | 0 | 41 | 0 | 35 | 0 | 30 | 0 | 26 | 0 | 14 | 0 | 7 | 0 | 7 | 0 |
| 92 to 121 | 464 | 53684 | 94 | 2 | 46 | 0 | 32 | 0 | 25 | 0 | 12 | 0 | 11 | 0 | 10 | 0 | 6 | 0 | 2 | 0 | 2 | 0 |
| 122 (100%) | 338 | 42 | 27 | 0 | 7 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

## VIII. Threats to validity

We list (and classify following [48]) threats to the validity of our work here:

- *Construct Validity:* A possible threat is the validity of the metrics used to reach our conclusions. The results would be different if instead of counting the number of inputs that fail with at least one metamorphic rule, we required multiple failures for each input to be counted. They would also look different depending on the metamorphic relations we take into account, or if the modified images are out of scope of the system under test. The image aggregation metric can also change the results depending on whether we measure the mean, median, minimum or maximum error of all keypoints in an output.
  To mitigate this, we consider the median error as an image aggregation metric, which is less sensitive to outliers compared to the mean or the maximum. We also show the number of failed inputs for different selections of metamorphic relations. We choose to count the inputs when they fail for only one metamorphic relation instead of many because this is enough to show that the framework found a problem that would be relevant to the application engineers. Finally, we show how a wide and varied selection of different metamorphic relations do not completely subsume each other, so the errors found come from the test with various different rules, and not only one rule finding all the problems.
  We also note that in this work we are not trying to find specific problems in the SUT, and so the influence of the metrics discussed here is reduced. They show that MET-POSE can be used to find problems in such systems, but the focus is not on which specific problems we found.
- *Internal Validity:* Another threat is that the results shown were found by chance or due to mistakes in our implementation, and not because of the causes that we expect. To mitigate this, we have carefully inspected our implementation, and manually confirmed each step done until the results aggregation. We have also added an *identity* metamorphic relation as a sanity check to ensure that the system results do not change when we run the system with the same input twice, in which case we would need to run our experiments a number of times to reach a statistically significant conclusion.
- *External Validity:* The current work may not generalise to other datasets or systems beyond the ones in this work. It is also not explored how effective it would be when used in a real application setting with a specific set of relevant input features and output requirements. In general, DL models are evolving rapidly and so it is common that many methods are rendered obsolete quickly.
  To mitigate these, we have designed this framework in a way that it is highly dataset and system agnostic: any image is processed equally, not depending on its colour, number or type of subjects, etc.; and MET-POSE tests the SUT as a black-box. The only change required to use different ground truth annotations, or different systems, is due to the non-standardised pose estimation outputs and annotations, which is unavoidable. We have also left not only the specific metamorphic relations but also the error metric open to modification for different applications, facilitating adaptability to different contexts.

## IX. Conclusions and Future Work

In this work we propose MET-POSE, a metamorphic testing-based framework for pose estimation systems. We have focused on making this framework both system- and dataset-agnostic, due to the rapid developments of the area of pose estimation systems. This should help the approach remain relevant even if architectures or methodological changes occur in the area.

We applied MET-POSE to *Mediapipe Holistic* using two datasets from different application domains. Results show that MET-POSE can detect failures of the system without ground truth data, i.e., without human annotations. They also show that, by leaving the definition of the metamorphic rules and the metrics used to assess violation of these rules open to the user, MET-POSE is adaptable and can test pose estimation systems for different use cases.

As future work, we plan to apply MET-POSE to different systems and use cases. This will first confirm the framework's adaptability. We also plan to show that, by using application-specific rules, practitioners who build these pose estimation systems can gain a better understanding of how they operate and even use this information towards repairing their systems.

## REFERENCES


[1] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *MIA*, vol. 42, pp. 60–88, 2017.

[2] R. Holmes, E. Rushe, M. De Coster, M. Bonnaerens, S. Satoh, A. Sugimoto, and A. Ventresque, "From scarcity to understanding: Transfer learning for the extremely low resource irish sign language," in *ICCV*, 2023.

[3] Z. Martin, S. Hendricks, and A. Patel, "Automated tackle injury risk assessment in contact-based sports-a rugby union example," in *CVPR*, 2021.

[4] I. Grishchenko and V. Bazarevsky, "MediaPipe Holistic — Simultaneous Face, Hand and Pose Prediction, on Device," https://research.google/blog/mediapipe-holistic-simultaneous-face-hand-and-pose-prediction-on-device/, [Accessed 19-08-2024].

[5] D. C. Luvizon, D. Picard, and H. Tabia, "2d/3d pose estimation and action recognition using multitask deep learning," in *CVPR*, 2018.

[6] "Mediapipe pose landmark detection system," [Accessed 21-01-2025]. [Online]. Available: https://ai.google.dev/edge/mediapipe/solutions/vision/pose_landmarker

[7] A. Birhane, S. Dehdashtian, V. Prabhu, and V. Boddeti, "The dark side of dataset scaling: Evaluating racial classification in multimodal models," in *FAccT*, 2024.

[8] A. Birhane, V. U. Prabhu, and J. Whaley, "Auditing saliency cropping algorithms," in *WACV*, 2022.

[9] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh *et al.*, "Ethical and social risks of harm from language models," *arXiv preprint arXiv:2112.04359*, 2021.

[10] I. D. Raji, T. Gebru, M. Mitchell, J. Buolamwini, J. Lee, and E. Denton, "Saving face: Investigating the ethical concerns of facial recognition auditing," in *AIES*, 2020.

[11] N. Hoehing, E. Rushe, and A. Ventresque, "What's left can't be right–the remaining positional incompetence of contrastive vision-language models," *arXiv preprint arXiv:2311.11477*, 2023.

[12] F. Liu, G. Emerson, and N. Collier, "Visual spatial reasoning," *TACL*, vol. 11, 2023.

[13] E. T. Barr, M. Harman, P. McMinn, M. Shahbaz, and S. Yoo, "The oracle problem in software testing: A survey," *TOSEM*, vol. 41, no. 5, 2014.

[14] L. Ma, F. Zhang, J. Sun, M. Xue, B. Li, F. Juefei-Xu, C. Xie, L. Li, Y. Liu, J. Zhao *et al.*, "Deepmutation: Mutation testing of deep learning systems," in *ISSRE*. IEEE, 2018.

[15] A. Dwarakanath, M. Ahuja, S. Sikand, R. M. Rao, R. J. C. Bose, N. Dubash, and S. Podder, "Identifying implementation bugs in machine learning based image classifiers using metamorphic testing," in *ISSTA*, 2018.

[16] L. Ma, F. Juefei-Xu, F. Zhang, J. Sun, M. Xue, B. Li, C. Chen, T. Su, L. Li, Y. Liu *et al.*, "Deepgauge: Multi-granularity testing criteria for deep learning systems," in *ASE*, 2018.

[17] M. Pu, C. Y. Chong, and M. K. Lim, "Robustness evaluation in hand pose estimation models using metamorphic testing," in *MET)*, 2023.

[18] X. Xie, J. W. Ho, C. Murphy, G. Kaiser, B. Xu, and T. Y. Chen, "Testing and validating machine learning classifiers by metamorphic testing," *JSS*, vol. 84, no. 4, 2011.

[19] "opencv implemmentation of gamma correction," [Accessed 21-01-2025]. [Online]. Available: https://docs.opencv.org/3.4/d3/dc1/tutorial_basic_linear_transform.html

[20] "opencv implemmentation of bilateral filtering," [Accessed 21-01-2025]. [Online]. Available: https://docs.opencv.org/4.x/d4/d13/tutorial_py_filtering.html

[21] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness," 2022. [Online]. Available: https://arxiv.org/abs/1811.12231

[22] M. De Coster, M. Van Herreweghe, and J. Dambre, "Sign language recognition with transformer networks," in *LREC*, 2020.

[23] R. Holmes, E. Rushe, F. Fowley, and A. Ventresque, "Improving signer independent sign language recognition for low resource languages," in *SLTAT*, 2022.

[24] "opencv implemmentation of hsv colorspace encoding," [Accessed 21-01-2025]. [Online]. Available: https://docs.opencv.org/3.4/df/d9d/tutorial_py_colorspaces.html

[25] I. Grishchenko, V. Bazarevsky, A. Zanfir, E. G. Bazavan, M. Zanfir, R. Yee, K. Raveendran, M. Zhdanovich, M. Grundmann, and C. Sminchisescu, "Blazepose ghum holistic: Real-time 3d human landmarks and pose estimation," *arXiv preprint arXiv:2206.11678*, 2022.

[26] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "Blazepose: On-device real-time body pose tracking," *arXiv preprint arXiv:2006.10204*, 2020.

[27] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," in *CVPR*, 2018.

[28] B. Sapp and B. Taskar, "Modec: Multimodal decomposable models for human pose estimation," in *CVPR*, 2013.

[29] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft coco: Common objects in context," *arXiv preprint arXiv:1405.0312*, 2014.

[30] G. A. for Developers, "Model Card: MediaPipe BlazePose GHUM 3D," https://storage.googleapis.com/mediapipe-assets/Model%20Card%20BlazePose%20GHUM%203D.pdf, 2021, [Online; Accessed 2 July, 2024].

[31] ——, "Model Card: MediaPipe FaceMesh," https://storage.googleapis.com/mediapipe-assets/Model%20Card%20MediaPipe%20Face%20Mesh%20V2.pdf, 2022, [Online; Accessed 25 September, 2024].

[32] ——, "Model Card: MediaPipe Hands (Lite/Full)," https://storage.googleapis.com/mediapipe-assets/Model%20Card%20Hand%20Tracking%20(Lite_Full)%20with%20Fairness%20Oct%202021.pdf, 2021, [Online; Accessed 25 September, 2024].

[33] M. Duran, T. Laurent, E. Rushe, and A. Ventresque, "Companion repository for "Metamorphic Testing for Pose Estimation Systems"," [Accessed 21-01-2025]. [Online]. Available: https://github.com/MatoFD/MeT-Pose

[34] M. Kassab, "Testing practices of software in safety critical systems: Industrial survey." in *ICEIS (2)*, 2018.

[35] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," 2019. [Online]. Available: https://arxiv.org/abs/1903.12261

[36] J. Mohapatra, T.-W. Weng, P.-Y. Chen, S. Liu, and L. Daniel, "Towards verifying robustness of neural networks against a family of semantic perturbations," in *CVPR*, 2020.

[37] N. Drenkow and M. Unberath, "Robustclevr: A benchmark and framework for evaluating robustness in object-centric learning," in *WACV*, 2024.

[38] N. Humbatova, G. Jahangirova, G. Bavota, V. Riccio, A. Stocco, and P. Tonella, "Taxonomy of real faults in deep learning systems," in *ICSE*, 2020.

[39] J. Kim, R. Feldt, and S. Yoo, "Evaluating surprise adequacy for deep learning system testing," *ACM TOSEM*, vol. 32, no. 2, 2023.

[40] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *FAccT*. PMLR, 2018.

[41] T. Laurent, P. Arcaini, X.-Y. Zhang, and F. Ishikawa, "Metamorphic testing of an autonomous delivery robots scheduler," in *ICST*. IEEE, 2024.

[42] P. Naidu, H. Gudaparthi, and N. Niu, "Metamorphic testing for convolutional neural networks: Relations over image classification," in *IRI*. IEEE, 2021.

[43] S. Wu, Y. Hu, Y. Wang, J. Gu, J. Meng, L. Fan, Z. Luan, X. Wang, and Y. Zhou, "Combating missed recalls in e-commerce search: A cot-prompting testing approach," in *FSE*, ser. FSE 2024. New York, NY, USA: Association for Computing Machinery, 2024. [Online]. Available: https://doi.org/10.1145/3663529.3663842

[44] X. Wang, G. Yi, and Y. Wang, "Automated functional testing of search engines using metamorphic testing," in *QRS*, 2021.

[45] J. Bozic and F. Wotawa, "Testing chatbots using metamorphic relations," in *Testing Software and Systems*, C. Gaston, N. Kosmatov, and P. Le Gall, Eds. Cham: Springer International Publishing, 2019.

[46] S. Helge, B. Nassim, G. Arnaud, and L. Nadjib, "Evaluating human trajectory prediction with metamorphic testing," *arXiv preprint arXiv:2407.18756*, 2024.

[47] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in *FAccT*, 2019.

[48] R. Feldt and A. Magazinius, "Validity threats in empirical software engineering research - an initial survey," in *SEKE*, 2010.