

TCG-AR: Real-Time Multi-View Augmented Reality for Trading Card Game Streaming

Anthony Cioppa^{*}, Antoine Verdonck^{*}, Maxim Henry,
Marc Van Droogenbroeck, and Raphaël La Rocca

University of Liège, Belgium anthony.cioppa@uliege.be

Abstract. Trading card games are increasingly played and broadcast online, yet live streams remain mostly limited to flat top-down footage of the playing area. Augmenting such streams with virtual models of the played cards would improve the viewing experience, but most existing systems rely on instrumented playing surfaces and embedded chips, which are costly and impractical for casual players and large-scale events. In this work, we present TCG-AR, a novel real-time pipeline that augments trading card games using ordinary RGB cameras alone, without any physical markers or specialized hardware. Our pipeline detects, orients, and identifies the cards on the board, renders virtual content onto each card across all views, and can additionally compose a broadcast-style view that summarizes the game state for spectators, streaming the augmented feeds to standard broadcasting software such as OBS. To train the detection, orientation, and identification models without manual labeling, we introduce an automatic procedure that generates annotated synthetic training data from a reference set of card images. Then, we evaluate several trained models on a new manually annotated dataset with real images, analyzing performance and runtime throughput that determine real-world usability. Overall, by relying only on commodity cameras and hardware, and by open-sourcing all code, models, and datasets, this work aims to serve as a reference for real-time trading card recognition and to make real-time augmented-reality streaming accessible to the broader community of players and streamers.

Keywords: Augmented reality · Object detection · Metric learning · Real-time systems · Trading card games

1 Introduction

Trading card games, such as the Pokémon, Yu-Gi-Oh!, Lorcana, One Piece, and Magic: The Gathering are played by millions of people worldwide and have developed a large online presence [3, 54]. Their cultural reach continues to grow, with 2026 marking the thirtieth anniversary of both the Pokémon franchise and its trading card game, drawing renewed public attention worldwide [60]. In the current social-media era, players increasingly record and broadcast their games

^{*} Equal contributions. Code: <https://github.com/ULiege-VIULab/tcg-ar>

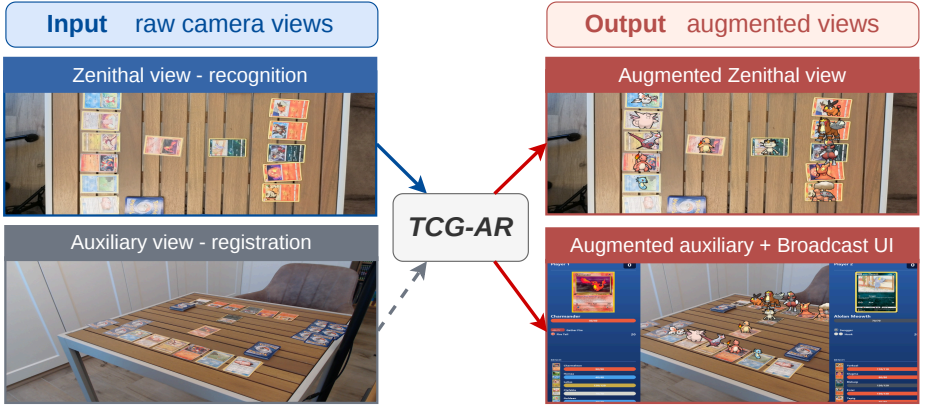


Fig. 1: Trading Card Game – Augmented Reality (TCG-AR). Given one or more camera views of a trading card game (left), TCG-AR detects, orients, and identifies the played cards, and renders a virtual model of the corresponding creature onto each card, in every view and in real time (right). The augmented feeds can be streamed directly to standard broadcasting software. No markers, chips or instrumented surfaces are required. Illustrated here on the Pokémon trading card game.

on platforms such as YouTube or Twitch, both for casual sharing and for competitive events [59]. However, these broadcasts remain visually limited: a stream typically shows a flat top-down view of the playing area, in which the cards appear as static pieces of printed cardboard, sometimes supplemented by additional views of the players. For viewers who are not already invested in the game, this presentation conveys little of the action and offers limited spectacle.

A natural way to enrich such broadcasts is through Augmented Reality (AR), which integrates virtual content into a view of the real world [1, 4]. Applied to a trading card game, AR can replace each played card on the stream with an animated model of the creature it represents, drawing on a long-standing direction in AR gaming that overlays virtual content onto physical play [61] to bring the broadcast closer to the experience depicted in the franchise’s associated media. A popular illustration of this idea was demonstrated by the content creator *Super-Zouloux* [56], whose system displays three-dimensional monster models on top of Yu-Gi-Oh! cards during a live game. While the result is visually convincing, it is achieved through dedicated hardware: a chip is embedded in every card, and the playing surface is instrumented with readers that detect which card is present at each position. Such a setup is costly and laborious, as it requires chipping all cards of both players, building a specialized board, and configuring the system, which places it out of reach of casual players and impractical for events that broadcast many games in parallel. Other systems that augment card games with virtual models also rely on instrumented surfaces [28] or markers attached to the cards [29] to locate and recognize them. More recently, consumer tools have used a webcam or phone camera to identify a single trading card for cataloging [14]

or broadcast overlays [23, 25], but these target collection and pack-opening use cases where a card is presented alone and unobstructed, rather than during an actual game where cards are laid out together on a board.

In this work, we address this gap by proposing TCG-AR, a novel real-time pipeline that augments trading card games using ordinary RGB cameras alone, without any physical markers or specialized hardware. Our pipeline, illustrated in Fig. 1, ingests one or more camera feeds and creates a game state by detecting, orienting, and identifying the cards from a zenithal view. The predicted card positions are then mapped into each additional view by registration, allowing a virtual model to be rendered onto every card across all feeds. From the game state, our pipeline also composes a broadcast-style view that summarizes the match for spectators. The augmented feeds are then streamed to standard broadcasting software such as OBS, integrating directly into the tools content-creators already use. To keep the stream fluid despite the computational cost of recognizing the game state, our pipeline separates a fast rendering path, which produces the augmented videos at a high frame rate, from the slower recognition path, which periodically updates the game state. We design the system to be agnostic to specific trading card games, using the Pokémon trading card game only as our demonstration case. To do so, we train the detection, orientation, and identification models on synthetic data produced by our automatic generation procedure, avoiding any manual annotation, and evaluate the performance on a collected and manually annotated dataset of real card games. At run time, our game state recognition path reliably detects, orients, and identifies each card, while the rendering path runs at a high frame rate to place an animated model on every card and stream the augmented feeds to standard broadcasting software over Real Time Streaming Protocol (RTSP). Beyond its technical design, our aim is also practical: by relying only on commodity cameras and consumer-grade gaming hardware, we seek to make real-time AR streaming accessible to the broader community of trading card game players and streamers.

Contributions. We summarize our contributions as follows: (i) We present TCG-AR, a real-time pipeline that detects and identifies trading cards from an RGB camera during real gameplay and renders augmented-reality content onto the corresponding cards across one or more views, without any physical markers or instrumented playing surface. TCG-AR further composes a broadcast-style view of the game state and streams the operator-selected feeds to standard broadcasting software via RTSP. (ii) We propose an automatic, game-agnostic procedure for generating synthetic training data from a reference set of card scans, producing detection, orientation, and identification datasets with annotations obtained at no manual labeling cost, which allows direct adaptation of TCG-AR to any trading card game. (iii) We collect and manually annotate a real evaluation dataset of the Pokémon trading card game, captured under varied conditions, to assess the game state recognition Sim2Real gap. (iv) We report a quantitative and qualitative study of our pipeline, analyzing detection and identification performance, comparing several model variants for detection, orientation, and identification, and characterizing the streaming runtime throughput.

2 Related Works

2.1 Augmented reality

Augmented reality aims to embed virtual content within a view of the real world, in real time and spatially registered [1], nowadays extended to modern tracking, display, and interaction technologies [4]. Placing virtual content correctly in a moving-camera setting requires continuously estimating the camera pose relative to the scene, the problem at the core of vision-based AR registration [37]. Classical solutions track the camera geometrically from natural image features [40], while learning-based methods regress the camera pose directly from a single RGB image [26, 27]. When the scene is planar and the cameras are fixed, placing overlays across views reduces to estimating a homography between them, classically by detecting and matching keypoints and fitting the transformation robustly with RANSAC [8, 15]. In this case, hand-crafted features such as SIFT [34] and ORB [47] remain widely used, while learned detectors and matchers, including SuperPoint [11], SuperGlue [48], LightGlue [32], and LoFTR [55], improve matching under low texture and large viewpoint change. Tabletop and game scenarios have been a recurring target for AR. For example, Rematas *et al.* [43] reconstruct a soccer game from ordinary video to replay it on a tabletop. For trading card games, early systems rely on markers or sensor-equipped surfaces. ARTable projects content onto an instrumented table [28], and popular hobbyist systems achieve convincing overlays by embedding chips in the cards and reading them with dedicated hardware [56]. More recent consumer tools instead recognize cards directly from RGB images, identifying a single card held up to the camera to catalogue it [14] or to overlay its name, image, and market price on a stream [23, 25]; these operate on one isolated card at a time and neither localize multiple cards on a board nor register the overlay across viewpoints. Unlike these, our TCG-AR pipeline recognizes the whole board during play, including overlaid and obstructed cards, from RGB images alone.

Object detection. Locating objects in an image is nowadays dominated by deep learning detectors, split between two-stage methods that propose and then classify regions of interest, from R-CNN [18] to Faster R-CNN [44], and one-stage methods that predict boxes directly, such as YOLO [42], SSD [33], and RetinaNet with its focal loss [31], alongside efficiency-oriented [58] and transformer-based set-prediction designs [6]. When objects are not axis-aligned, axis-aligned boxes discard their orientation, motivating oriented object detection, studied for aerial and text imagery through rotated-anchor and refinement methods [20, 71], transformer-based oriented detectors [36], and the Oriented R-CNN detector we adopt, which generates oriented proposals efficiently [69]; a compact classifier on an efficient backbone [57] that resolves the residual 180° orientation ambiguity. Training detectors without manual annotation is commonly achieved with synthetic data: domain randomization varies lighting, texture, and pose so that models trained on non-photorealistic renders transfer to real images [63, 64], cut-and-paste synthesis composites object crops onto diverse backgrounds with

automatically derived labels [13, 17], and large rendered datasets serve driving and scene understanding [16, 45, 46]. Applied to cards, these ideas have produced convolutional detectors for cluttered poker scenes [7], card classifiers [38], and automatically generated card datasets for bridge broadcasting [68]. In this work, we detect cards with an oriented detector trained on synthetic data, and evaluate it on real footage.

Identification and re-identification. Distinguishing instances within a single object category, and recognizing the same instance again across frames or cameras, is a fine-grained problem that is typically cast as metric learning rather than flat classification. Images are mapped to an embedding in which a nearest-neighbor search retrieves the closest reference. This embedding-and-retrieval paradigm was popularized by Siamese [5, 73] and contrastive [19] formulations. Such paradigms rely on losses such as the triplet, which produces a compact descriptor compared by Euclidean distance [49], and was subsequently refined through better losses and sampling of informative pairs and triplets [51, 65, 67], proxy- and classification-based objectives [10, 39, 74], and surveyed comparisons [41]. The same formulation underlies object re-identification, where an instance must be matched across views despite changes in appearance [22, 52, 53, 72], a setting well suited to many classes with few samples each. A trading card game is an extreme such case: tens of thousands of visually similar cards, with new ones released over time, which a fixed-output classifier cannot accommodate. For example, the Pokémon trading card game currently comprises up to 20,360 distinct cards, orders of magnitude more than a standard deck of playing cards (52). In this work, we therefore identify cards by retrieval against per-card reference embeddings, so the label set can grow without retraining, and reuse the same descriptors to re-identify cards in the stream.

3 Methodology

3.1 Pipeline overview and design constraints

TCG-AR takes as input one or more RGB camera feeds of a trading card game played on a flat surface, and produces the augmented feeds with a virtual model rendered on top of each played card. We formulate the task as game state recognition, defined as the set of cards currently on the board together with their position, orientation (Sec. 3.2), and identity (Sec. 3.3). Game state recognition is performed on a single zenithal view, where the cards appear with a near-frontal geometry. The remaining cameras serve only for output: they display the augmented result from additional viewpoints. The card positions estimated in the zenithal view are mapped into each auxiliary view through image registration (Sec. 3.4), where the virtual models are then drawn (Sec. 3.5).

Our TCG-AR pipeline, illustrated in Fig. 2, runs two concurrent paths so that the displayed frame rate is not tied to the cost of game state recognition. The recognition path updates the game state by performing detection, orientation, and identification of the cards on the zenithal view. The rendering path

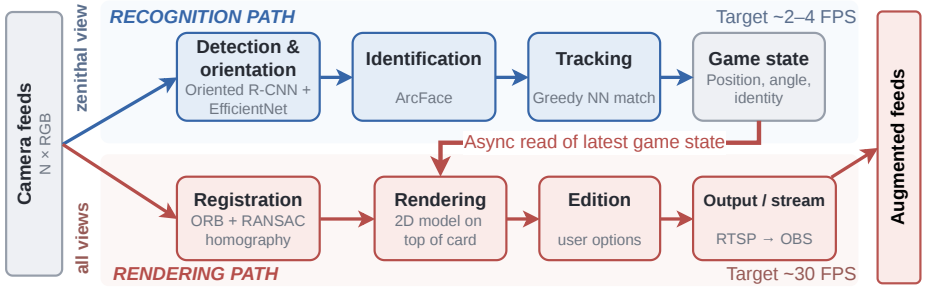


Fig. 2: Overview of our TCG-AR pipeline. TCG-AR runs two concurrent paths. The game state recognition path (top) operates on a zenithal view alone: it detects and orients the cards, then identifies and tracks them to estimate the game state, *i.e.*, the set of cards on the board with their identity, position, and orientation. The rendering path (bottom) operates on all views: it registers the auxiliary views to the zenithal reference, renders a virtual model onto each card from the current game state, applies user-selected edition options, and streams the augmented feeds to standard broadcasting software. The rendering path reads the latest game state asynchronously, so the displayed video stays fluid even when recognition is slower.

acquires frames, renders the current game state onto them, and streams the result, reading the most recent state asynchronously so the video stays fluid even when recognition is slower. The two paths face different timing requirements. The rendering path drives what the viewer sees and targets the 30 frames per second common to streaming setups. The recognition path is bound only by latency, since the board typically changes every several seconds: the output must reflect a change quickly enough to go unnoticed. Using the average human Simple Reaction Time (SRT) of roughly 250 ms as a reference [66], we target a game state refresh of about 4 frames per second as comfortable. This budget leads the recognition path to favor models and matching strategies that run on commodity hardware over heavier, more accurate alternatives; we motivate these choices per stage and report the resulting frame rates in Sec. 5. While our figures and experiments use the Pokémon trading card game, TCG-AR assumes only a reference set of card images, and applies to other trading card games without modification.

3.2 Card detection and orientation

The first stage of the game state recognition path aims to locate the cards in the zenithal view. While cards are rectangular and rigid, making them well suited to detection, they may not always be axis-aligned: players place and rotate cards freely, and in several trading card games, a rotation may encode a change in a card’s status. An axis-aligned bounding box would therefore discard the orientation of each card, whereas the subsequent identification stage (Sec. 3.3) needs an upright, canonical crop to compare a card against the reference set. We consequently detect cards with an oriented object detector, which outputs for each card a rotated bounding box parameterized by its center, width, height,

and in-plane angle. We compared several detectors and chose the Oriented R-CNN [69] model for this stage, a two-stage detector that generates oriented region proposals and refines them into rotated boxes. We adopt a ResNet-50 [21] backbone with a feature pyramid [30], rather than a heavier backbone or a transformer-based detector, because this configuration meets the timing budget while retaining sufficient performance (see Sec. 5). The detector is trained for a single object class, “card”, since this stage only needs to separate cards from the background and the playing surface.

However, while the detector localizes each card, it only resolves its orientation up to a sign: the rotated box fixes the card’s axis but not which of its two ends points up, leaving a 180° ambiguity. This ambiguity matters because the identification stage expects an upright one. We resolve this orientation task with a lightweight orientation classifier applied to each detected card crop, which predicts whether the crop is upright or rotated by 180° . Crops predicted as inverted are then flipped before identification. We investigated several architectures and chose the EfficientNet-B0 network [57] whose classification head is replaced by a two-way output. Together, the two components turn the zenithal frame into a set of upright, canonically oriented card crops with known positions, which the identification stage can use to determine each card’s identity.

3.3 Card identification

Given the upright card crops produced by the detection and orientation stage, the identification stage assigns the identity of the card it depicts to each crop, drawn from the reference set of all cards in the game. Trading card games may contain on the order of tens of thousands of distinct cards, many of which differ only in small regions such as a name, a numeric value, or an illustration. Furthermore, the set of cards grows over time as new ones are released every 3 to 6 months. Hence, training a flat classifier over so many near-identical classes, and retraining it whenever the set changes, is impractical. We therefore cast identification as a metric learning task, following the embedding-and-retrieval approach popularized for face recognition [49]. A model maps first each card crop to a fixed-length descriptor, and then a card is identified by comparing its descriptor to a set of reference descriptors, one per known card, computed once from the reference card images. Identification reduces to a nearest-neighbor search in descriptor space, so adding or removing cards only requires inserting or deleting reference descriptors, with no retraining. We use a ResNet-50 [21] backbone, from which the descriptor is produced by one of two heads that we compare empirically. The first head follows the original triplet formulation [49]: a fully connected layer maps the backbone features to a 128-dimensional descriptor trained with a triplet margin loss. Each triplet consists of a reference card image as anchor, a second view of the same card as positive, and a different card as negative, and the loss drives same-card descriptors closer together than different-card descriptors by at least a fixed margin under the Euclidean distance. The second head, which we adopt by default, replaces the triplet objective with the additive angular margin loss of ArcFace [10]: each card defines a

class whose 512-dimensional embedding is constrained to a hypersphere, an angular margin is enforced between classes during training, and cards are matched at test time by cosine similarity. This optimizes the embedding against all cards jointly rather than from sampled triplets, which we find yields a more robust descriptor on real crops. Both heads share the same backbone, the same one-per-card reference descriptors, and the same nearest-neighbor retrieval to ensure fair comparison. Since the ArcFace head defines one class per card and needs only the reference images, whereas selecting difficult negatives for the triplet head [19] uses game-specific metadata (detailed for Pokémon in the supplementary material), deploying ArcFace keeps both the data generation and the identification model agnostic to the specific trading card game.

Comparing each crop against the descriptors of all cards in the game is a source of avoidable errors, since most of those cards cannot appear in a given match. A game is typically played from decks that together contain only a small fraction of the existing cards. We can exploit this by restricting the nearest-neighbor search to the cards declared in the players’ deck lists, which reduces the candidate set from the full reference set to a few tens/hundreds of cards. This restriction removes the possibility of matching to a card that is not in play, which improves identification accuracy. Finally, since identification is run per frame, but the cards on the board persist across frames, recomputing every identity at every recognition step is wasteful and prone to flicker. Therefore, we associate the cards detected in the current frame with those in the previous game state by a greedy nearest-position matching to improve identification stability.

3.4 Multi-view registration

To render a model onto a card in an auxiliary view, the card’s location in that view must be known. Rather than running detection and identification separately in every view, we recover the card locations in the auxiliary views by mapping the zenithal positions into them. Since cards lie on a common planar surface and the cameras are static, each auxiliary view is related to the zenithal view by a homography, *i.e.*, a 3×3 projective transformation between the two image planes. We estimate each homography once, at initialization, from a pair of frames showing the same board. The automatic procedure first detects key-points in both frames with ORB [47]. From correspondences, the homography is fitted with RANSAC [15]. When the automatic procedure fails to find enough correspondences, for instance under repetitive or low-texture backgrounds, we fall back to a manual step in which the user provides the corresponding pairs.

3.5 Rendering, edition, and streaming output

The rendering path turns the current game state into the augmented video. For each frame, and for each card present in the game state, it draws a virtual model of the entity the card represents on top of each card. Virtual content is drawn from a model database that pairs each card identity with a renderable asset. We render two-dimensional assets, compositing a flat, animated sprite onto the

Table 1: Composition of our datasets. The three synthetic datasets are generated automatically from the reference card images and annotated without human intervention; the real dataset is captured and manually annotated, and used only for evaluation. K denotes the number of cards in the reference set ($K = 20,360$ for Pokémon).

Dataset	Purpose	# Images	Resolution	Annotation
Synthetic detection	train/val/test	18,000 / 1,000 / 1,000	1920 × 1080	oriented box, card
Synthetic orientation	train/val/test	18,000 / 1,000 / 1,000	224 × 224	upright / 180°
Synthetic identification	train/test	5 K / K	224 × 224	card identity
Real	evaluation	51 (832 inst.)	1920 × 1080	oriented box + identity

card region as well as broadcast-style information about the game, such as the current active card, the bench, as well as card information (*e.g.*, type, health, attacks, etc.). Further details are provided in supplementary material. Finally, the output stage makes the augmented feeds available to external software. Each enabled feed is emitted as a Real-Time Streaming Protocol (RTSP) stream, in up to three forms per camera, namely its raw image, its augmented version, and its broadcast composition, with only the operator-selected forms produced and streamed. Rather than provide editing and broadcasting tools of its own, the system delegates these to existing streaming software: the RTSP streams can be ingested directly by common tools such as OBS, where the operator composes the final broadcast. This keeps TCG-AR focused on producing the augmented views and lets it fit into the workflows players and streamers already use.

4 Datasets

The detection, orientation, and identification models are trained on synthetic data that we generate automatically, with no manual annotation. Our generation procedure takes as its only game-specific input a reference set of card images, one canonical image per card, which is readily available for trading card games [2, 50]. From this reference set, our procedure generates a separate datasets for each task, with the annotations derived directly from how each image is constructed. We instantiate it on the Pokémon trading card game, whose reference set contains $K = 20,360$ different cards at the time of writing, spanning 1,025 distinct creatures and many more trainer, energy, and object cards. Since nothing in the procedure is specific to a particular game beyond the reference images, the same generator applies to other trading card games. We additionally collect a real, manually annotated dataset, used only for evaluation, to measure how well the resulting models transfer to genuine play. Figure 3 shows examples of both synthetic and real images, and Tab. 1 summarizes their composition.

4.1 Synthetic data generation

The detection dataset is built by compositing cards onto background images. We draw a background photograph from a collection of $B = 28,074$ textures gath-

ered from public sources [9, 75], resize it to the camera resolution of 1920×1080 , and sample between 5 and 15 cards to place on it. For each card, we select a reference image at random, apply a random scale and in-plane rotation, and paste it at a random location, allowing partial overlaps so that the model sees crowded boards. We automatically record the card annotation as its four corner coordinates with a single “card” class label. We then apply randomized augmentations, each with a fixed probability, to narrow the gap to real footage: localized lighting effects that brighten parts of the scene and introduce saturated highlights, imitating spotlights, the striped pattern of neon tubes, and the specular glare that glossy cards produce ($p = 0.33$), global saturation changes ($p = 0.10$), additive and impulse noise ($p = 0.20$), and a perspective warp that displaces the image corners within a margin ($p = 0.20$). The dataset comprises 18,000 training, 1,000 validation, and 1,000 test generated images.

The orientation dataset is produced by the same compositing process restricted to a single card per image, rendered at the classifier’s input resolution of 224×224 . Each card is placed close to upright, within $\pm 10^\circ$, and half of the crops are then rotated by 180° ; the applied rotation automatically determines the binary label. This yields a balanced set of upright and inverted crops, 20,000 in total, including a validation and test split.

The identification dataset provides, for each card in the reference set, several views of that card under varied appearance. For every reference card we generate six 224×224 crops using the augmentations described above, so that the network sees each card under different lighting, color, and noise while its identity is fixed. Of the six crops per card, five are used for training and one for validation, giving 5K training and K validation crops, with 122,160 images in total.

4.2 Real evaluation dataset

Synthetic data lets us train without manual annotation, but it cannot fully reproduce a real game: physical cards under real lighting, captured by consumer cameras, with the imperfections the recognition path must tolerate in deployment. To measure how well the models transfer to these conditions, that is, the Sim2Real gap, we collect and manually annotate a dataset of real trading card scenes, including sleeved cards and holofoil prints. This dataset is used only for evaluation; no model is trained or fine-tuned on it. We capture real scenes of the Pokémon trading card game with a consumer webcam (Insta 360 Link 2C) mounted zenithally above the board, recording at a resolution of 1920×1080 . The scenes are captured under varied real lighting and deliberately include crowded boards, partially occluded cards, and face-down cards, so that the evaluation reflects realistic rather than idealized play. In total, the dataset contains 51 images with 832 manually annotated card instances. Each instance is annotated manually with an oriented bounding box, given as four corner coordinates, and the identity of the card, expressed using the same reference identifiers as the card database. The annotations were produced by the authors with a custom annotation tool and cross-checked.



Fig. 3: Examples from our datasets. (a) Synthetic training data generated automatically from the reference card images: a detection sample with multiple cards composited onto a textured background, single-card crops for orientation, and identification crops of one card under varied lighting, color, and noise. (b) Real evaluation scenes captured under different lighting and board conditions, manually annotated with oriented boxes and card identities. Illustrated for the Pokémon trading card game.

5 Experiments

We evaluate our TCG-AR pipeline along two axes: the performance of the recognition stages, and the runtime throughput of the pipeline, the latter determining whether the system is usable for live streaming. All experiments are run on a single machine equipped with an AMD Ryzen 5 3600X CPU and an NVIDIA RTX 2080 Ti GPU, which corresponds to a consumer-grade streaming setup. We report performance on both the synthetic test data and on the real evaluation dataset, which together quantify the Sim2Real gap, and we report throughput across the recognition stages and the output forms that drive it. To justify our architectural choices, we further compare five oriented detectors, four orientation backbones, and two identification heads, as reported in Tab. 2. All details on model training parameters can be found in supplementary material.

5.1 Detection and orientation

Regarding detection, we compare five oriented detectors, all with a ResNet-50 backbone and trained under identical data and training parameters, so the comparison isolates the architecture. On real data, all five models reach essentially the same mAP of about 0.90 (Tab. 2), which confirms that card detection transfers robustly from synthetic to real regardless of the detector. The architectures separate mainly on the synthetic test set, where the two-stage Oriented R-CNN is clearly strongest (0.91 mAP, against 0.82 for the one-stage RetinaNet and FCOS and lower for the heavier RoI Transformer and Gliding Vertex), and on speed, where FCOS is fastest on the GPU and the one-stage detectors are fastest

Table 2: Game state recognition quantitative results. Five oriented detectors, four orientation backbones, and the two identification heads are evaluated on both the synthetic and real test sets, with reported per-stage speed (frames per second on a board of 14 cards). *Real (deck)* restricts the identification candidate set to the 631 unique cards present in the real data; identification speed is reported as full / deck. The deployed pipeline uses Oriented R-CNN, EfficientNet-B0, and ArcFace.

	Synthetic	Real	Real (deck)	FPS _{GPU}	FPS _{CPU}
<i>Detection</i>	<i>mAP / Recall</i>				
Oriented R-CNN [69]	0.91 / 0.91	0.90 / 0.90	–	14.0	0.6
Rotated RetinaNet [31]	0.82 / 0.83	0.91 / 0.92	–	14.1	0.9
Rotated FCOS [62]	0.82 / 0.81	0.90 / 0.91	–	15.9	0.9
RoI Transformer [12]	0.73 / 0.75	0.91 / 0.90	–	13.1	0.4
Gliding Vertex [70]	0.64 / 0.68	0.90 / 0.90	–	14.7	0.4
<i>Orientation</i>	<i>Accuracy</i>				
EfficientNet-B0 [57]	100.0	98.7	–	119	2
ResNet-18 [21]	99.9	98.9	–	144	4
MobileNetV3-S [24]	100.0	98.7	–	203	11
ShuffleNetV2 [35]	100.0	98.1	–	108	7
<i>Identification</i>	<i>Top-1 accuracy / Top-5 accuracy</i>				
Triplet [49]	48.5 / 71.7	1.7 / 4.4	17.3 / 36.8	57 / 60	1.3 / 1.3
ArcFace [10]	97.0 / 99.5	85.1 / 96.2	96.4 / 99.2	53 / 55	1.2 / 1.2

on the CPU. We keep Oriented R-CNN in the deployed pipeline for its balance of accuracy and speed (0.90 real mAP at 14 FPS on the GPU). Regarding orientation, the 180° classifier is a simpler binary task: all four backbones reach $\geq 98\%$ accuracy on real crops and essentially 100% on synthetic (Tab. 2).

5.2 Identification

Table 2 shows that the choice of descriptor head is decisive on real data. The triplet head, although adequate on synthetic crops, collapses on real cards, reaching only 1.7% top-1 against the full reference set, whereas the ArcFace head reaches 85.1% under the same open-set protocol. The candidate-set size is the second lever. Without restriction, each crop is matched against the full reference set of $K = 20,360$ cards; with the deck-list restriction, the candidate set shrinks to the 631 unique cards. On the real dataset, this raises ArcFace top-1 accuracy from 85.1% to 96.4% by removing cards that cannot appear. Its effect on identification speed is small, however (Tab. 2): the per-frame cost is dominated by the embedding network rather than the nearest-neighbor search, which is cheap even over the full reference set on the GPU. Finally, the current models proved robust across the lighting conditions we captured.

Table 3: Rendering throughput, measured on the evaluation machine on one 1920×1080 view and a busy board of 14 cards. *Composite* is the cost of drawing the virtual characters, *resize + copy* the cost of preparing the preview and the shared-memory frame, and *total* their sum. The broadcast form’s side panels are cached, so a warm frame (cache hit) is much cheaper than a cold one (panel rebuild). *Streams@30 FPS* is how many such views fit in the 40ms display budget. Each form is produced only when selected, so producing fewer forms frees both rendering and encoder budget.

Output form	Composite (ms)	Resize + copy (ms)	Total (ms)	View-FPS	Streams @ 30 FPS
Raw	0.0	1.9	1.9	532	17
Augmented (AR)	6.6	1.9	8.5	118	3
Broadcast (warm)	27.4	1.9	29.3	34	1
Broadcast (cold)	35.9	1.9	37.7	27	< 1
All three forms	34.0	5.6	39.7	25	< 1

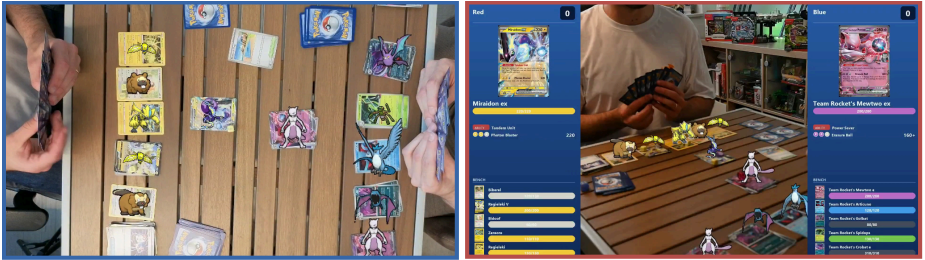


Fig. 4: Qualitative results. Augmented output produced by TCG-AR. Left: the zenithal view after augmentation, with a virtual model rendered onto each recognized card. Right: Auxiliary view augmented from the same game state through registration with side panels showing card information. Several cards carry damage counters and status markers that occlude part of their design. TCG-AR identifies them correctly.

5.3 Rendering and output

The rendering path runs concurrently with recognition. Table 3 reports its cost per output form on a single view. Raw output is essentially free; the augmented form composites the sprites at 118 view-FPS, and the broadcast form, which additionally composites the side panels, runs at 34 view-FPS; producing all three forms of a view together still sustains 25 FPS. Meanwhile the recognition path refreshes the game state at around 10 FPS on the GPU.

5.4 Qualitative results

Figure 4 shows the zenithal view before and after augmentation together with auxiliary views rendered from the same game state and the broadcast composition. Because cards are augmented during play rather than presented in isolation, they are frequently partially covered: damage counters, status markers, and neighboring cards placed on top may occlude parts of the artwork and text.

Our identification stage still retrieves the correct card in these cases, matching it from the visible remainder rather than from any single region of the card. Furthermore, we observed stable behavior across the lighting conditions we captured, in line with the robustness of the recognition stages reported above; the residual errors are those inherited from recognition, namely occasional misses on heavily cluttered boards and confusions between visually near-identical cards.

5.5 Discussion and limitations

Taken together, the results show that a markerless RGB-only pipeline can augment a trading card game in real time, but they also delineate where it currently falls short. Detection now transfers well from synthetic to real data, and the ArcFace descriptor makes identification robust on real crops (Tab. 2); the deck-list restriction further improves identification when deck lists are available. The remaining components are deliberately simple and limit the system at the edges: registration relies on a planar homography between fixed cameras and its automatic estimation is the least reliable stage in practice, while temporal matching compares only consecutive states and therefore loses cards under occlusion or a passing hand; these would benefit from more robust feature matching and a short temporal voting window, respectively. Finally, the rendering path is restricted to two-dimensional sprites; extending it to three-dimensional models, and validating it on trading card games beyond the one studied here, are the natural next steps toward the accessible streaming tool this work aims to provide.

6 Conclusion

We presented TCG-AR, a real-time pipeline that augments trading card games from ordinary RGB cameras, without the chips, or instrumented surfaces that previous systems require. From a single zenithal view, the pipeline detects, orients, and identifies the played cards to estimate the game state, and renders a virtual model onto each card across one or more registered views, and can compose a broadcast-style summary of the match, streaming the result to standard broadcasting software. To train the recognition models without manual labeling, we introduced an automatic, game-agnostic procedure that generates annotated synthetic data from a reference set of card images, and we evaluated the trained models on a new manually annotated real dataset, measuring how they transfer to genuine play. The design is organized around a real-time budget, separating a fast rendering path from a slower recognition path, so that the augmented video stays fluid while the game state is refreshed quickly enough for live viewing. By relying only on commodity hardware and releasing our code, models, and datasets, we hope to bring augmented-reality streaming within reach of the broad community of trading card game players and streamers, and to provide a foundation for further work on vision-based augmentation of these games.

Bibliography

- [1] Azuma, R.T.: A survey of augmented reality. Presence: Teleoperators Virtual Environ. **6**(4), 355–385 (Aug 1997). <https://doi.org/10.1162/pres.1997.6.4.355>
- [2] Backes, A.: Pokémon TCG SDK for Python. Github site, <https://github.com/PokemonTCG/pokemon-tcg-sdk-python> (2021)
- [3] Big Orbit Cards: A brief history of trading card games. <https://www.bigorbitcards.co.uk/blog/a-brief-history-of-trading-card-games.html> (2023)
- [4] Billinghurst, M., Clark, A., Lee, G.: A survey of augmented reality. Found. Trends® Human-Comput. Interact. **8**(2-3), 73–272 (Mar 2015). <https://doi.org/10.1561/11000000049>
- [5] Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a “siamese” time delay neural network. In: Adv. Neural Inf. Process. Syst. (NeurIPS). vol. 6, pp. 737–744 (1993)
- [6] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Eur. Conf. Comput. Vis. (ECCV). Lect. Notes Comput. Sci., vol. 12346, pp. 213–229. Springer Int. Publ. (2020). https://doi.org/10.1007/978-3-030-58452-8_13
- [7] Chen, Q., Rigall, E., Wang, X., Fan, H., Dong, J.: Poker watcher: Playing card detection based on EfficientDet and sandglass block. In: Int. Conf. Aware. Sci. Technol. (icast). pp. 1–6. IEEE, Qingdao, China (Dec 2020). <https://doi.org/10.1109/icast51195.2020.9319468>
- [8] Chum, O., Pajdla, T., Sturm, P.: The geometric error for homographies. Comput. Vis. Image Underst. **97**(1), 86–102 (Jan 2005). <https://doi.org/10.1016/j.cviu.2004.03.004>
- [9] Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 3606–3613. IEEE, Columbus, OH, USA (Jun 2014). <https://doi.org/10.1109/cvpr.2014.461>
- [10] Deng, J., Guo, J., Yang, J., Xue, N., Kotsia, I., Zafeiriou, S.: ArcFace: Additive angular margin loss for deep face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **44**(10), 5962–5979 (Oct 2022). <https://doi.org/10.1109/tpami.2021.3087709>
- [11] DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperPoint: Self-supervised interest point detection and description. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW). pp. 337–33712. IEEE, Salt Lake City, UT, USA (Jun 2018). <https://doi.org/10.1109/cvprw.2018.00060>
- [12] Ding, J., Xue, N., Long, Y., Xia, G.S., Lu, Q.: Learning RoI transformer for oriented object detection in aerial images. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 2844–2853. IEEE, Long Beach, CA, USA (Jun 2019). <https://doi.org/10.1109/cvpr.2019.00296>

- [13] Dwibedi, D., Misra, I., Hebert, M.: Cut, paste and learn: Surprisingly easy synthesis for instance detection. In: IEEE Int. Conf. Comput. Vis. (ICCV). pp. 1310–1319. IEEE, Venice, Italy (Oct 2017). <https://doi.org/10.1109/iccv.2017.146>
- [14] Eyevo: Eyevo: Pokémon TCG scanner. Web site
- [15] Fischler, M.A., Bolles, R.C.: Random sample consensus. Commun. ACM **24**(6), 381–395 (Jun 1981). <https://doi.org/10.1145/358669.358692>
- [16] Fonder, M., Van Droogenbroeck, M.: Mid-air: A multi-modal dataset for extremely low altitude drone flights. In: IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), UAVision. pp. 553–562. IEEE, Long Beach, CA, USA (Jun 2019). <https://doi.org/10.1109/cvprw.2019.00081>
- [17] Georgakis, G., Mousavian, A., Berg, A., Kosecka, J.: Synthesizing training data for object detection in indoor scenes. In: Robotics: Science and Systems XIII. vol. XIII, pp. 1–9. Cambridge, MA, USA (Jul 2017). <https://doi.org/10.15607/rss.2017.xiii.043>
- [18] Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 580–587. Columbus, OH, USA (Jun 2014). <https://doi.org/10.1109/CVPR.2014.81>
- [19] Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). vol. 2, pp. 1735–1742. Inst. Electr. Electron. Eng. (IEEE), New York City, NY, USA (Jun 2006). <https://doi.org/10.1109/CVPR.2006.100>
- [20] Han, J., Ding, J., Xue, N., Xia, G.S.: ReDet: A rotation-equivariant detector for aerial object detection. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 2785–2794. IEEE, Nashville, TN, USA (Jun 2021). <https://doi.org/10.1109/cvpr46437.2021.00281>
- [21] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 770–778. IEEE, Las Vegas, NV, USA (Jun 2016). <https://doi.org/10.1109/CVPR.2016.90>
- [22] He, S., Luo, H., Wang, P., Wang, F., Li, H., Jiang, W.: TransReID: Transformer-based object re-identification. In: IEEE/CVF Int. Conf. Comput. Vis. (ICCV). pp. 14993–15002. IEEE, Montréal, Can. (Oct 2021). <https://doi.org/10.1109/iccv48922.2021.01474>
- [23] HitScanTCG: HitScanTCG: AI-powered card scanning for streamers. Web site, <https://www.hitscantcg.com> (2025)
- [24] Howard, A., Sandler, M., Chen, B., Wang, W., Chen, L.C., Tan, M., Chu, G., Vasudevan, V., Zhu, Y., Pang, R., Adam, H., Le, Q.: Searching for MobileNetV3. In: IEEE/CVF Int. Conf. Comput. Vis. (ICCV). pp. 1314–1324. IEEE, Seoul, South Korea (Oct 2019). <https://doi.org/10.1109/iccv.2019.00140>
- [25] Hype Overlay: Hype Overlay: OBS-native trading card game scanner. Web site, <https://hypeoverlay.com> (2026)

- [26] Kendall, A., Cipolla, R.: Modelling uncertainty in deep learning for camera relocalization. In: IEEE Int. Conf. Robot. Autom. (ICRA). pp. 4762–4769. IEEE, Stockholm, Sweden (May 2016). <https://doi.org/10.1109/icra.2016.7487679>
- [27] Kendall, A., Grimes, M., Cipolla, R.: PoseNet: A convolutional network for real-time 6-DOF camera relocalization. In: IEEE Int. Conf. Comput. Vis. (ICCV). pp. 2938–2946. Inst. Electr. Electron. Eng. (IEEE), Santiago, Chile (Dec 2015). <https://doi.org/10.1109/iccv.2015.336>
- [28] Lam, A.H.T., Chow, K.C.H., Yau, E.H.H., Lyu, M.R.: ART: augmented reality table for interactive trading card game. In: ACM Int. Conf. Virtual Real. Contin. Its Appl. pp. 357–360. ACM, Hong Kong, China (Jun 2006). <https://doi.org/10.1145/1128923.1128987>
- [29] Lee, W., Woo, W., Lee, J.: TARBoard: Tangible augmented reality system for table-top game environment. In: PerGames. pp. 1–5. Munich, Germany (May 2005)
- [30] Lin, T.Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 2117–2125. Honolulu, HI, USA (Jul 2017). <https://doi.org/10.1109/CVPR.2017.106>
- [31] Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection. arXiv **abs/1708.02002** (2017)
- [32] Lindenberger, P., Sarlin, P.E., Pollefeys, M.: LightGlue: Local feature matching at light speed. In: IEEE/CVF Int. Conf. Comput. Vis. (ICCV). pp. 17581–17592. IEEE, Paris, Fr. (Oct 2023). <https://doi.org/10.1109/iccv51070.2023.01616>
- [33] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.: SSD: Single shot multibox detector. arXiv **abs/1512.02325** (2016)
- [34] Lowe, D.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (Nov 2004)
- [35] Ma, N., Zhang, X., Zheng, H.T., Sun, J.: ShuffleNet v2: Practical guidelines for efficient CNN architecture design. In: Eur. Conf. Comput. Vis. (ECCV). Lect. Notes Comput. Sci., vol. 11218, pp. 122–138. Springer Int. Publ. (2018). https://doi.org/10.1007/978-3-030-01264-9_8
- [36] Ma, T., Mao, M., Zheng, H., Gao, P., Wang, X., Han, S., Ding, E., Zhang, B., Doermann, D.: Oriented object detection with transformer. arXiv **abs/2106.03146** (2021). <https://doi.org/10.48550/arXiv.2106.03146>
- [37] Marchand, E., Uchiyama, H., Spindler, F.: Pose estimation for augmented reality: A hands-on survey. IEEE Trans. Vis. Comput. Graph. **22**(12), 2633–2651 (Dec 2016). <https://doi.org/10.1109/tvcg.2015.2513408>
- [38] Mittal, K., Gill, K.S., Chauhan, R., Sharma, M., Sunil, G.: Playing cards classification and detection using sequential CNN model through machine learning techniques using artificial intelligence. In: Int. Conf. E-mobility, Power Control Smart Syst. (ICEMPS). pp. 1–4. IEEE, Thiruvananthapuram, India (Apr 2024). <https://doi.org/10.1109/icemps60684.2024.10559365>

- [39] Movshovitz-Attias, Y., Toshev, A., Leung, T.K., Ioffe, S., Singh, S.: No fuss distance metric learning using proxies. In: IEEE Int. Conf. Comput. Vis. (ICCV). pp. 360–368. IEEE, Venice, Italy (Oct 2017). <https://doi.org/10.1109/iccv.2017.47>
- [40] Mur-Artal, R., Montiel, J.M.M., Tardos, J.D.: ORB-SLAM: A versatile and accurate monocular SLAM system. IEEE Trans. Robot. **31**(5), 1147–1163 (Oct 2015). <https://doi.org/10.1109/tro.2015.2463671>
- [41] Musgrave, K., Belongie, S., Lim, S.N.: A metric learning reality check. In: Eur. Conf. Comput. Vis. (ECCV). Lect. Notes Comput. Sci., vol. 12370, pp. 681–699. Springer Int. Publ. (2020). https://doi.org/10.1007/978-3-030-58595-2_41
- [42] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 779–788. Inst. Electr. Electron. Eng. (IEEE), Las Vegas, NV, USA (Jun 2016). <https://doi.org/10.1109/cvpr.2016.91>
- [43] Rematas, K., Kemelmacher-Shlizerman, I., Curless, B., Seitz, S.: Soccer on your tabletop. In: IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 4738–4747. Salt Lake City, UT, USA (Jun 2018). <https://doi.org/10.1109/CVPR.2018.00498>
- [44] Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (Jun 2017). <https://doi.org/10.1109/TPAMI.2016.2577031>
- [45] Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: Eur. Conf. Comput. Vis. (ECCV). Lect. Notes Comput. Sci., vol. 9906, pp. 102–118. Springer Int. Publ. (2016). https://doi.org/10.1007/978-3-319-46475-6_7
- [46] Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 3234–3243. IEEE, Las Vegas, NV, USA (Jun 2016). <https://doi.org/10.1109/cvpr.2016.352>
- [47] Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: ORB: An efficient alternative to SIFT or SURF. In: IEEE Int. Conf. Comput. Vis. (ICCV). pp. 2564–2571. IEEE, Barcelona, Spain (Nov 2011). <https://doi.org/10.1109/iccv.2011.6126544>
- [48] Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperGlue: Learning feature matching with graph neural networks. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 4937–4946. IEEE, Seattle, WA, USA (Jun 2020). <https://doi.org/10.1109/CVPR42600.2020.00499>
- [49] Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: A unified embedding for face recognition and clustering. In: IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 815–823. IEEE, Boston, MA, USA (Jun 2015). <https://doi.org/10.1109/cvpr.2015.7298682>

- [50] Scrydex: Scrydex: TCG API and toolkit for pokémon, magic, lorcan, and yu-gi-oh! Web site, <https://scrydex.com> (2025)
- [51] Sohn, K.: Improved deep metric learning with multi-class N-pair loss objective. In: Adv. Neural Inf. Process. Syst. (NeurIPS). vol. 29, pp. 1857–1865. Curran Assoc. Inc., Barcelona, Spain (Dec 2016)
- [52] Somers, V.: Person Re-Identification and its Application to Multi-Object Tracking. Ph.D. thesis, Cathol. Univ. Louvain, Belg. (May 1995)
- [53] Somers, V., De Vleeschouwer, C., Alahi, A.: Body part-based representation learning for occluded person re-identification. In: IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV). pp. 1613–1623. IEEE, Waikoloa, HI, USA (Jan 2023). <https://doi.org/10.1109/wacv56688.2023.00166>
- [54] Statista: Pokémon trading card game - statistics & facts. <https://www.statista.com/statistics/672617/most-valuable-pokemon-trading-cards/> (2025)
- [55] Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: LoFTR: Detector-free local feature matching with transformers. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 8918–8927. IEEE, Nashville, TN, USA (Jun 2021). <https://doi.org/10.1109/cvpr46437.2021.00881>
- [56] SuperZouloux: Le rêve de tous les fans de Yu-Gi-Oh! YouTube video, <https://www.youtube.com/watch?v=64-LfbggqKI> (Oct 2005), <https://www.youtube.com/watch?v=64-LfbggqKI>
- [57] Tan, M., Le, Q.V.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: Int. Conf. Mach. Learn. (ICML). vol. 97, pp. 6105–6114. PMLR, Long Beach, CA, USA (Jun 2019)
- [58] Tan, M., Pang, R., Le, Q.V.: EfficientDet: Scalable and efficient object detection. In: IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 10778–10787. Seattle, WA, USA (Jun 2020). <https://doi.org/10.1109/CVPR42600.2020.01079>
- [59] The Pokémon Company International: 2025 pokémon world championships. <https://worlds.2025.pokemon.com/en-us/championships/> (2025)
- [60] The Pokémon Company International: Media alert: Pokémon unveils new pokémon trading card game: 30th celebration expansion, commemorating 30 years. <https://press.pokemon.com/en/releases/MEDIA-ALERT-Pokemon-Unveils-New-Pokemon-Trading-Card-Game-30th-Celebra> (2026)
- [61] Thomas, B.H.: A survey of visual, mixed, and augmented reality gaming. Comput. Entertain. **10**(1), 1–33 (Oct 2012). <https://doi.org/10.1145/2381876.2381879>
- [62] Tian, Z., Shen, C., Chen, H., He, T.: FCOS: A simple and strong anchor-free object detector. IEEE Trans. Pattern Anal. Mach. Intell. **44**(4), 1922–1933 (Apr 2020). <https://doi.org/10.1109/tpami.2020.3032166>
- [63] 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: IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS). pp. 23–30. IEEE, Vancouver, Can. (Sept 2017). <https://doi.org/10.1109/iros.2017.8202133>

- [64] Tremblay, J., Prakash, A., Acuna, D., Brophy, M., Jampani, V., Anil, C., To, T., Cameracci, E., Boochoon, S., Birchfield, S.: Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW). pp. 1082–10828. IEEE, Salt Lake City, UT, USA (Jun 2018). <https://doi.org/10.1109/cvprw.2018.00143>
- [65] Wang, X., Han, X., Huang, W., Dong, D., Scott, M.R.: Multi-similarity loss with general pair weighting for deep metric learning. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 5017–5025. IEEE, Long Beach, CA, USA (Jun 2019). <https://doi.org/10.1109/cvpr.2019.00516>
- [66] Woods, D.L., Wyma, J.M., Yund, E.W., Herron, T.J., Reed, B.: Factors influencing the latency of simple reaction time. *Front. Hum. Neurosci.* **9** (Mar 2015). <https://doi.org/10.3389/fnhum.2015.00131>
- [67] Wu, C.Y., Manmatha, R., Smola, A.J., Krahenbuhl, P.: Sampling matters in deep embedding learning. In: IEEE Int. Conf. Comput. Vis. (ICCV). pp. 2859–2867. IEEE, Venice, Italy (Oct 2017). <https://doi.org/10.1109/iccv.2017.309>
- [68] Wzorek, P., Kryjak, T.: Training dataset generation for bridge game registration. *arXiv abs/2109.11861* (2021). <https://doi.org/10.48550/arXiv.2109.11861>
- [69] Xie, X., Cheng, G., Wang, J., Yao, X., Han, J.: Oriented r-CNN for object detection. In: IEEE/CVF Int. Conf. Comput. Vis. (ICCV). pp. 3500–3509. IEEE, Montréal, Can. (Oct 2021). <https://doi.org/10.1109/iccv48922.2021.00350>
- [70] Xu, Y., Fu, M., Wang, Q., Wang, Y., Chen, K., Xia, G.S., Bai, X.: Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**(4), 1452–1459 (Apr 2021). <https://doi.org/10.1109/tpami.2020.2974745>
- [71] Yang, X., Yan, J., Feng, Z., He, T.: R3Det: Refined single-stage detector with feature refinement for rotating object. In: AAAI Conf. Artif. Intell. vol. 35, pp. 3163–3171. Assoc. Adv. Artif. Intell. (AAAI) (May 2021). <https://doi.org/10.1609/aaai.v35i4.16426>
- [72] Ye, M., Chen, S., Li, C., Zheng, W.S., Crandall, D., Du, B.: Transformer for object re-identification: A survey. *Int. J. Comput. Vis.* **133**(5), 2410–2440 (Nov 2024). <https://doi.org/10.1007/s11263-024-02284-4>
- [73] Yi, D., Lei, Z., Liao, S., Li, S.Z.: Deep metric learning for person re-identification. In: Int. Conf. Pattern Recognit. (ICPR). pp. 34–39. IEEE, Stockholm, Sweden (Aug 2014). <https://doi.org/10.1109/icpr.2014.16>
- [74] Zhai, A., Wu, H.Y.: Classification is a strong baseline for deep metric learning. *arXiv abs/1811.12649* (2018). <https://doi.org/10.48550/arXiv.1811.12649>
- [75] Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ADE20K dataset. In: IEEE Conf. Comput. Vis. Pattern

Recognit. (CVPR). pp. 5122–5130. IEEE, Honolulu, HI, USA (Jul 2017).
<https://doi.org/10.1109/cvpr.2017.544>

7 Supplementary Material

This appendix presents more information about the interface and card-rendering implementation, training details for the detection, orientation, and identification models, failure cases, and generalization.

7.1 Interface and card rendering

On launch, the interface lists the detected cameras, lets the operator designate the zenithal reference view, and runs the registration; when the automatic homography is unsatisfactory the operator can pick corresponding key points by hand to refine it. Each feed advertises its streaming address, so the augmented and broadcast outputs can be picked up directly by standard broadcasting software.

The sprite is drawn at an adjustable scale, set by default to twice the card size so that the model is clearly visible, and when several models overlap in a view they are composited back-to-front, ordered by their projected vertical position so that nearer models occlude farther ones. To convey the opposition between the two sides of the board, the models on one side are mirrored horizontally so that the two players' creatures face each other; since this is decided from the board position, the same orientation is applied consistently across all views. Cards that do not depict a creature, such as trainer and energy cards, are recognized and listed but left un-augmented, as they have no associated model. The same entity may admit several visual variants, which the renderer exposes as selectable options, although such variants are specific to the game being augmented rather than to the method. On the zenithal view the sprite is centred on its card, whereas on the auxiliary (side) views it is anchored by its base and rendered slightly smaller, so that the creature appears to stand on the card from an oblique viewpoint.

Beyond augmenting individual feeds, the rendering path can assemble a broadcast-style view that presents the game state as a television production would, illustrated in Fig. 4. A chosen feed is placed at the center, and a panel on each side summarizes one player's position: the active creature is shown with its card illustration, name, and game information, namely its health, type, and attacks with their energy costs, while the creatures on the bench appear below as a row of up to five cards. The cards in each panel are taken directly from the recognized game state, the active and benched cards being inferred from their positions on the board, whereas the information the system cannot observe, namely the players' names, scores, the venue, and each active creature's health and status, is entered by the operator through the interface described next. Because the panels depend only on the game state and not on the viewpoint, their rendered layer is computed once and shared across all views, and rebuilt only when the game state or the operator-entered information changes, which keeps the composition within the real-time budget. The composed view is streamed like any other output. A recognized-cards side panel lists every detected card

with its reference information and exposes the per-card rendering options (visual variant, scale, and whether a given card is augmented), illustrated in Fig. 5. A settings dialog collects machine-level parameters such as the streaming server path.

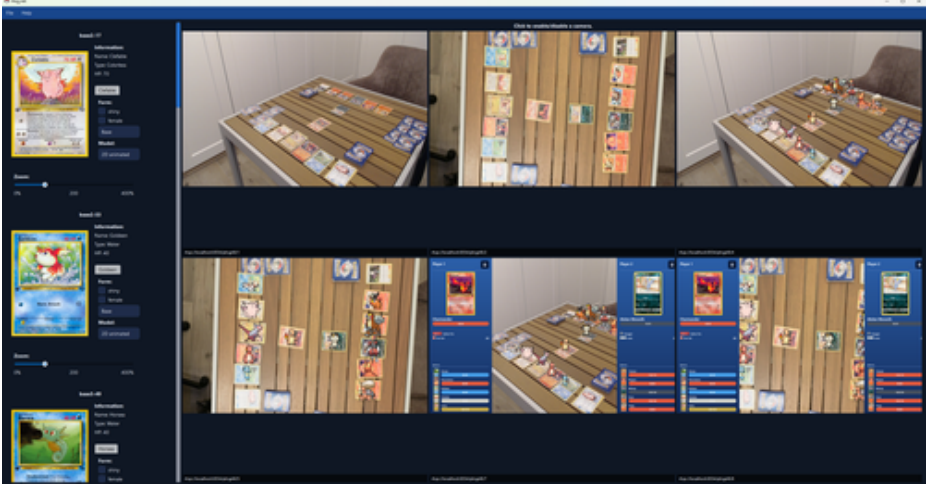


Fig. 5: GUI composition produced by TCG-AR. The user can control which camera feeds to send to OBS. A side panel provides extra card annotation and control over the model size and forms, *e.g.*, shiny or male versus female versions of the creature.

Furthermore, a small user interface lets the operator control the augmented output without editing code. It displays the available feeds in a grid, showing for each camera its raw, augmented, and broadcast forms side by side, each independently selectable so that only the chosen feeds are produced and streamed, which saves computation. A side panel lists the cards currently recognized, with their reference information, and exposes the per-card rendering options described above. A separate panel collects the broadcast information the system cannot observe, namely the player names, scores, venue, and each active creature’s current health and status, which feeds the broadcast composition above, and the interface also exposes minor conveniences such as displaying the streaming address of each feed. This interface is intended to keep the system usable by streamers without technical expertise, in line with our goal of accessibility.

Beyond the broadcast controls described above, the interface exposes the rest of the pipeline so that a streamer can operate it without editing code; Fig. 6 shows the deck-selection panel. A setup panel presents the full card database as a searchable grid (by name, by number, and by set) from which the operator assembles a deck for each of the two players; a card may be placed in both decks, and identification is then restricted to the union of the two decks. Decks can be named, saved and reloaded across sessions, and the operator chooses whether to

use the pre-trained model as-is on the restricted candidate set or to fine-tune it on the deck before starting.

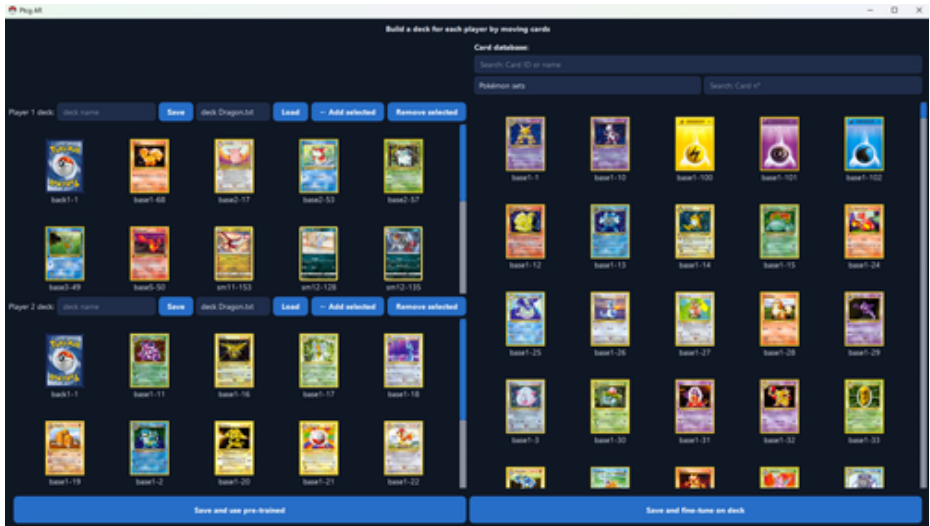


Fig. 6: Operator interface (deck-selection panel). The full card database is shown as a searchable grid (right) from which a deck is assembled for each player (left); a card may belong to both decks, and identification is restricted to their union. Decks can be named, saved and reloaded, and the operator chooses between using the pre-trained model as-is or fine-tuning it on the deck. Shown for the Pokémon trading card game.

7.2 Implementation and training details

This section reports the full training and evaluation setup for every model (detection, orientation, identification), so that all results in the main paper can be reproduced. Unless stated otherwise, all models are trained only on the synthetic training split and evaluated on the held-out synthetic test split and on the real set; no real data is used for training. After removing wrong scans (Sec. 7.2) and adding the card back as one extra identity, the reference set contains $K=20,360$ cards.

Hardware and software All experiments run on a single machine: AMD Ryzen 5 3600X CPU and one NVIDIA RTX 2080 Ti GPU. Software: Python 3.11, PyTorch 2.0.1 / torchvision 0.15.2 (CUDA 11.8); the detector uses mmrotate 0.3.4, mmdet 2.28.2 and mmcv-full 1.7.2. Image pre-processing uses OpenCV and Pillow. All backbones are initialised from ImageNet weights.

Card detection The deployed detector is Oriented R-CNN [69] with a ResNet-50 [21] backbone (ImageNet-pretrained, first stage frozen, batch-norm statistics frozen) and a feature pyramid [30], a single object class (**card**), and the **1e90** angle convention. The four comparison detectors (Rotated RetinaNet [31], Rotated FCOS [62], RoI Transformer [12], Gliding Vertex [70]) reuse the *exact* data pipeline, optimiser and schedule below; only the detector architecture changes (each is the stock mmrotate **r50_fpn_1x_dota_1e90** model with **num_classes** set to 1). Images are resized to 1024×1024 (**RResize**); random rotated flips are applied horizontally, vertically and diagonally (each with probability 0.25); pixels are normalised with mean (123.675, 116.28, 103.53) and std (58.395, 57.12, 57.375) (RGB), then padded to a multiple of 32. SGD with learning rate 5×10^{-3} , momentum 0.9, weight decay 10^{-4} , gradient-norm clipping at 35. A linear warm-up over 500 iterations (ratio 1/3) precedes a step schedule that divides the learning rate by 10 at epochs 2 and 4. Training runs for 5 epochs with a batch size of 2 images on one GPU ($\sim 9,000$ iterations/epoch over the 18,000 training images); a checkpoint is saved every epoch. The RPN uses anchors of scale 8 with aspect ratios $\{0.5, 1, 2\}$ over strides $\{4, 8, 16, 32, 64\}$; the RoI head uses rotated RoI-align (output 7×7) and two shared fully connected layers of width 1024. At test time boxes are kept above a score of 0.8 with rotated NMS at IoU 0.5.

Card orientation The orientation classifier is a binary (upright vs. 180°) network on 224×224 crops. The deployed model is EfficientNet-B0 [57]; the comparison backbones are ResNet-18 [21], MobileNetV3-Small [24] and ShuffleNetV2 [35]. Each is ImageNet-pretrained with its classification head replaced by a two-way linear layer; everything else is identical so the comparison isolates the backbone. Resize to 224×224 , random horizontal flip ($p = 0.5$), the synthetic augmentation suite of Sec. 4.1 (lighting, saturation, noise, perspective), random sharpness adjustment ($p = 0.5$), then **ToTensor** and ImageNet normalisation (mean (0.485, 0.456, 0.406), std (0.229, 0.224, 0.225)). Validation/test use only resize and normalization. Adam with learning rate 10^{-3} , a step scheduler (decay every 5 epochs), cross-entropy loss, batch size 16, for 10 epochs; a checkpoint is saved every epoch.

Card identification Both identification heads share a ResNet-50 backbone (ImageNet-pretrained) on 224×224 inputs and are trained on the synthetic identification split (six augmented crops per card: five for training, one held out for validation). The train augmentation is: resize to 224×224 , Gaussian blur (kernel 5–9, $\sigma \in [0.1, 5]$), random sharpness adjustment ($p = 0.5$), **ToTensor**, ImageNet normalisation. At test time a card is identified by nearest neighbour to a set of reference descriptors, one per card, each computed once from the card’s reference image.

For triplet, the backbone’s final layer is replaced by $\text{Linear}(2048, 512) \rightarrow \text{ReLU} \rightarrow \text{Linear}(512, 128)$, giving a 128-d descriptor. Training uses the triplet margin loss (margin 1.0, Euclidean distance) with hard-negative mining (the

negative for each anchor is a visually close card rather than a random one), Adam at learning rate 10^{-3} with a step schedule (factor 1/3 every 4 epochs), batch size 16, for 20 epochs. Retrieval uses Euclidean distance.

For ArcFace, the final layer is replaced by $\text{BN}(2048) \rightarrow \text{Dropout}(0.5) \rightarrow \text{Linear}(2048, 512) \rightarrow \text{BN}(512)$, giving a 512-d embedding. Training uses an additive angular-margin softmax (ArcFace [10]) with scale $s = 30$ and margin $m = 0.50$, one output class per card. Optimisation is SGD with learning rate 10^{-3} , momentum 0.9, weight decay 5×10^{-4} , and a cosine-annealing schedule (over the 20 epochs, minimum learning rate $10^{-3} \times 10^{-3}$); cross-entropy on the margin logits; batch size 16. A checkpoint is saved every epoch. We found that the Adam/step recipe of the triplet head collapses the learning rate before the $\sim 20\text{k}$ -class margin head converges; the SGD/cosine recipe above is required. At test time the ArcFace classifier head is discarded and the 512-d embedding is compared by cosine similarity.

Hard negative sampling for identification The triplet head’s difficulty comes almost entirely from cards that look alike: different prints of the same creature, cards that share an artwork frame, or the many near-identical Trainer and Energy cards. Drawing the negative of each anchor uniformly at random would therefore spend most of the training budget on trivially easy pairs. Instead, we sample negatives with a category-aware curriculum driven by the card metadata, so that the loss concentrates on the confusions that actually occur. For a creature (Pokémon) anchor, with probability ≈ 0.5 the negative is another card of the *same species*, i.e., same national Pokédex number(s), type and subtype, a different print or illustration of the same creature, the hardest case (and we raise this probability when the anchor has no Pokédex number, such as a special art card); with probability ≈ 0.2 it is a card of the *same energy type* but a different species; with probability ≈ 0.1 it is any other creature; and otherwise the negative is drawn uniformly. Trainer and Energy anchors, which form large visually homogeneous families, draw their negative from their own category (80% of the time) before falling back to a random card. Figure 7 shows an anchor with one negative sampled from each tier. This sampling is used for the triplet head; the ArcFace head obtains an analogous effect implicitly, since its angular-margin softmax pushes every card away from all $\sim 20\text{k}$ class centers.

Finally, when deck lists are available, the descriptor network can additionally be fine-tuned for 10 epochs at learning rate 10^{-4} (triplet batch size 8; ArcFace resumed from the base weights) on the restricted candidate set before a session, further separating the cards that will actually be encountered. This is optional and is *not* used for the main-paper numbers, which evaluate the base models. Furthermore, it is interesting to note that our model generalizes well to cards written in other languages than English, such as French or German, who share similar illustrations.

Reference-set cleaning A small fraction of reference images are wrong scans whose stored picture is the card back. We embed every reference image with



Fig. 7: Hard-negative tiers used for the triplet head. For a creature anchor (left), negatives are drawn with decreasing probability from the same species (a different print of the same creature, *i.e.* the hardest case), the same energy type, and then any creature. Trainer and Energy anchors draw from their own category. The curriculum forces the descriptor to separate look-alike cards rather than easy random pairs. Shown for the Pokémon trading card game.

the base ArcFace model and flag those whose cosine similarity to the embedding of a known card back exceeds 0.85; flagged cards are excluded from training, retrieval and evaluation. The card back itself is retained as a separate identity so that face-down cards are recognized as such.

Evaluation protocols *Detection.* We report DOTA-style mean average precision (AP at IoU 0.5) and recall, on the synthetic test split (1,000 images) and the real set (51 images, 832 annotated instances); for the real set the test data paths are overridden to the real annotations, leaving everything else identical. Per-stage frame rate is measured by timing single-image inference over 100 images on the GPU and the CPU.

Orientation. We report top-1 accuracy of the binary upright/180° decision on the synthetic test split (1,000 crops) and on 636 crops derived from the real annotations (each ground-truth-upright real crop kept straight or rotated 180°, balanced). Forward frame rate is measured on the GPU and the CPU.

Identification. We report top-1 and top-5 accuracy by nearest-neighbour retrieval against the reference descriptors (cosine for ArcFace, Euclidean for triplet). Iden-

tification is evaluated on *ground-truth-upright* crops (orientation taken from the annotation), so that it is measured independently of the orientation model. On the synthetic test data each card’s held-out sixth crop is queried against the full reference set. On the real set (636 annotated crops) we report two regimes: the open set (all 20,360 cards as candidates) and the deck-restricted set (candidates limited to the 631 unique cards present in the real data).

Throughput. Per-stage recognition speed is reported as full zenithal frames per second assuming a busy board of 14 cards, on the GPU and the CPU; rendering cost is reported per output form (raw, augmented, broadcast).

7.3 Failure cases

The two recognition stages fail in different ways. The detectors behave as detectors of rectangular card-like regions: they are most reliable on well-separated, face-up cards and degrade on heavily cluttered arrangements, where they both miss cards and produce false positives, and where overlapping cards occasionally merge into a single box. Face-down cards are detected but, having no face, are identified as the card back.

The errors of the deployed ArcFace head are dominated by genuinely look-alike cards rather than by obvious mistakes. On the 636 real crops, the open set produces 95 top-1 errors, of which 71 (75%) still place the correct card within the top five retrieved references; only 24 are top-5 misses. Restricting the candidate set to the cards present in the decks removes most of these confusions, leaving 23 top-1 errors and only 5 top-5 misses. Figure 8 shows representative cases: the query is almost always retrieved alongside other prints of the same creature or cards sharing the same frame, and the few top-5 misses correspond to crops degraded by motion blur, glare or strong perspective rather than to a confusion between unrelated cards. This is the behaviour the hard-negative curriculum of Sec. 7.2 targets, and it explains why a small candidate restriction (the deck list) recovers most of the lost accuracy.

The registration is the least robust component. It is a simple planar mapping between fixed views performed at the beginning of the capture sequence. It is not a full six-degree-of-freedom camera pose estimation: it assumes the board is planar and the cameras do not move after initialization, which holds for a typical tabletop streaming setup but would not accommodate freely moving cameras. The automatic estimation is also the least reliable component of the pipeline in practice, mostly due to not enough distinctive keypoints in typical card game scenarios at the beginning of the game. The fallback manual annotation proves to be fast, with around 30 seconds required to annotate at least 4 common keypoints across views. In practice, since the cards themselves are richly textured, placing a few of them on the bare table before initialization gives enough distinctive keypoints to register the views. A more principled alternative, which we leave to future work, is to register directly from the cards rather than the background: each detected card supplies four coplanar corners, and since all cards lie on the table plane, corner correspondences between views determine the same planar homography without relying on background texture at all.

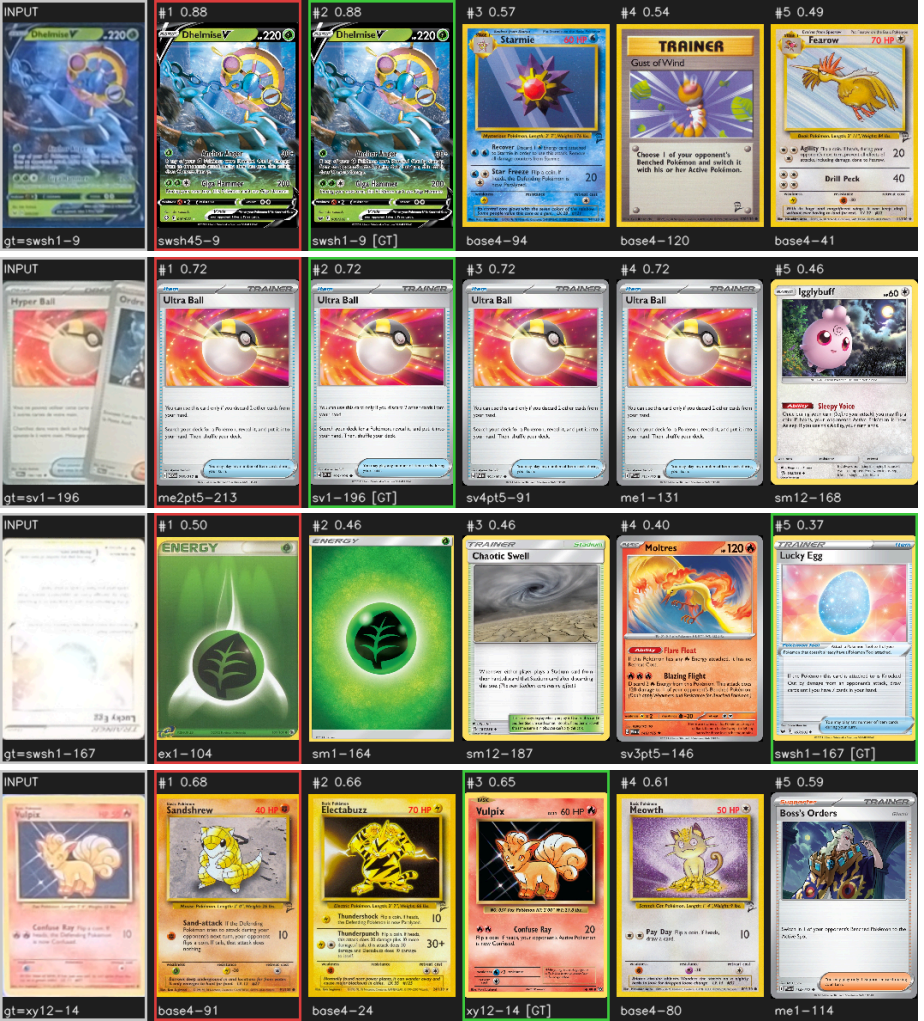


Fig. 8: Representative identification errors (real set, ArcFace). Each row shows a query crop (left) and the top-five retrieved reference cards; the ground-truth card, when present, is outlined. Most errors are confusions between near-identical cards, *e.g.*, different prints of the same creature or cards sharing an illustration frame, so the correct card usually remains within the top five.

7.4 Assessing generalization

A practical concern for a deployed system is whether it keeps working as new cards are released, since the trading card game grows continuously. Identification is designed for this: the ArcFace classifier head, which has one output per training card, is discarded at test time, and recognition is performed by nearest-neighbour retrieval against a per-card reference descriptor computed directly

from the card’s official image. A card the network never saw during training can therefore be recognised as soon as its reference image is added to the gallery, with no retraining. This is the property that lets the same models cover the full 20,360-card reference set while being trained on synthetic crops alone. A more direct protocol freezes the embedding trained on cards released up to a cutoff date and evaluates retrieval on the $>2,000$ cards released afterward, which the network has never seen; retrieval remains accurate on these unseen cards.

7.5 Scope of the recognized game state

The recognized game state is the set of visible, face-up cards, each with a position, an in-plane orientation, and an identity. A physical trading card game board carries more information than this, and we make the boundary explicit here. Several mechanics stack cards so that only the top one is visible: energy cards are placed beneath a Pokémon to power its attacks, and an evolution keeps its lower stages under the current one for the rest of the game. A zenithal camera sees only the top card, so attachments and the hidden stages of an evolution chain cannot be recovered from a single frame, and the pipeline treats what is beneath the top card as unobserved. Damage counters, status markers, and dice likewise convey numeric and categorical state that is not part of a card’s identity. The pipeline is robust to these as occluders, identifying a card from its visible remainder even when they cover part of it (Sec. 5.4), but it does not read their values; a creature’s remaining health or a poison or paralysis condition is therefore entered by the operator, and in-plane rotation is the only status signal read directly from the board.

Finally, tracking associates cards between consecutive game states by nearest-position matching, which stabilizes identities frame to frame but keeps no longer history. A card that becomes hidden, whether covered by a hand, another card, or a marker, leaves the state until it is visible again, at which point it is re-detected with no memory of its earlier identity. Persisting identities across occlusion, and reconstructing attachment and evolution relationships over the course of a game, would require a temporal model of the board and are left to future work. Energy attachment is a concrete case where these limits combine. A player attaches an energy from hand in a quick motion, often under a second, and tucks it under the Pokémon so that only a strip remains visible, with no standard placement across players. The recognition path refreshes at roughly four frames per second, so the placement and the hand passing over it can fall entirely between two recognition frames. Even once the board settles, the energy is mostly occluded and in no canonical orientation, so the detector and identifier may not recover it at all. Finally, an attachment is a relationship between two cards, “this energy powers that Pokémon,” whereas the recognized state is a flat set of cards with a position, orientation, and identity, and has no field to express it. The system therefore does not capture energy attachments: at best it reports the visible strip as a separate, possibly unidentified card, and more often the energy is absent from the state entirely. Representing such relationships, and sampling the board

densely enough to catch fast placements, would require both a temporal model and a relational state, and are left to future work.

7.6 Rendering assets and licensing

The rendering path pairs each recognized card identity with a renderable model drawn from an asset database. This database is not part of our system and is assembled at the user’s end from external sources; we redistribute none of its contents.

For the Pokémon demonstration, the database combines two kinds of external data. Card metadata and reference images, used both for identification and for the card panels of the broadcast view, are obtained through the Pokémon TCG API [2] and its associated card database.¹ The creature models composited onto the cards are two-dimensional animated sprites collected from a community sprite repository,² indexed against the National Pokédex ordering.³ Both are fetched by the user at setup time; the system stores only the mapping from card identity to asset, not the assets themselves.

The creatures, their sprites, and the card artwork are the intellectual property of the game’s publisher and are not openly licensed. Our release therefore comprises the pipeline, the trained detection, orientation, and identification models, and the synthetic-data generator, but no copyrighted card images or creature assets; each user retrieves those from the sources above under the terms that apply to them. The recognition models are trained on synthetically composited crops of the reference card images and do not embed redistributable copies of them, so the released models carry no card artwork. We note that the community sprite sources are themselves of unclear redistribution status, which is a further reason we treat them strictly as user-supplied inputs rather than shipping them; a deployment that requires clean licensing could substitute any asset set keyed by card identity, as the renderer is agnostic to the origin of the models.

Because the asset database is external and could be incomplete, the renderer defines an explicit fallback for cards it cannot illustrate. A card whose identity is recognized but for which the database holds no model, either a non-creature card such as a Trainer or Energy card or a creature absent from the current asset set, is detected, identified, and listed in the recognized-cards panel, but composited without an overlay. Augmentation of the remaining cards on the board is unaffected, so a missing asset degrades coverage gracefully rather than interrupting the stream.

¹ <https://dev.pokemontcg.io>

² https://projectpokemon.org/home/docs/spriteindex_148

³ https://www.pokepedia.fr/Liste_des_Pok%C3%A9mon_dans_l%27ordre_du_Pok%C3%A9dex_National