Hand-Shadow Poser

HAO XU* and YINQIAO WANG*, The Chinese University of Hong Kong, Hong Kong, China
NILOY J. MITRA, University College London, Adobe Research, United Kingdom
SHUAICHENG LIU, University of Electronic Science and Technology of China, China
PHENG-ANN HENG and CHI-WING FU, The Chinese University of Hong Kong, Hong Kong, China

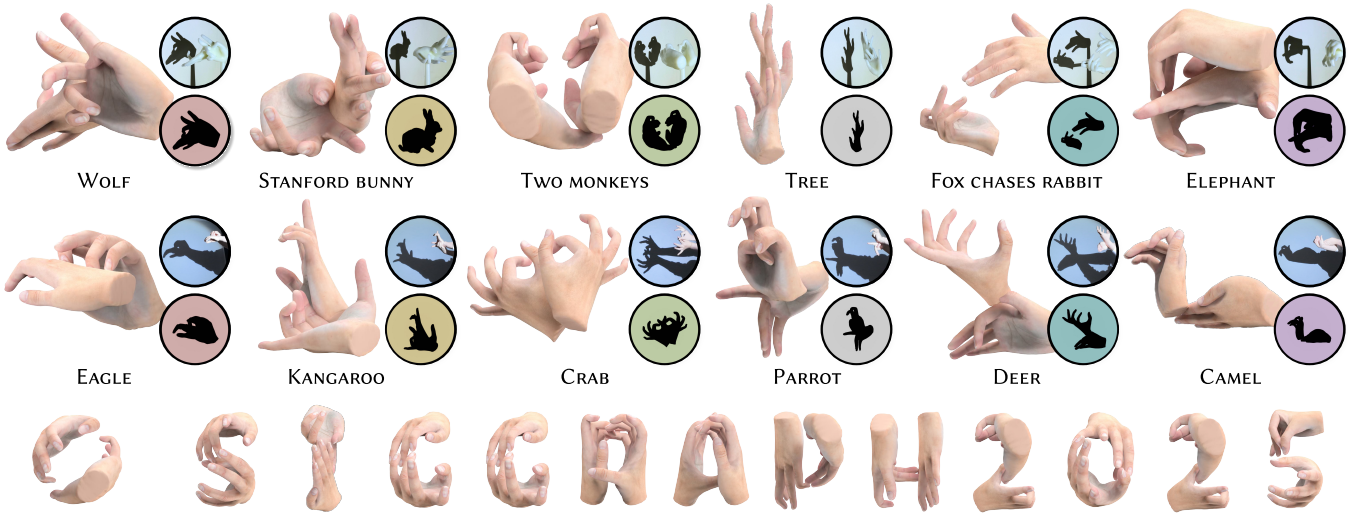


Fig. 1. 3D poses of left and right hands reconstructed by our method for producing shadows of different target objects. Lower insets show renderings of each 3D hand-pose result with the viewpoint set at the light source location, thus essentially revealing the final shadows produced by the respective hand poses. The upper insets in the first row show their 3D prints, whereas those in the second row show real shadows produced by human hands.

Hand shadow art is a captivating art form, creatively using hand shadows to reproduce expressive shapes on the wall. In this work, we study an inverse problem: given a target shape, find the poses of left and right hands that together best produce a shadow resembling the input. This problem is nontrivial, since the design space of 3D hand poses is huge while being restrictive due to anatomical constraints. Also, we need to attend to the input's shape and crucial features, though the input is colorless and textureless. To meet these challenges, we design Hand-Shadow Poser, a three-stage pipeline, to decouple the anatomical constraints (by hand) and semantic constraints (by shadow shape): (i) a generative hand assignment module to explore diverse but reasonable left/right-hand shape hypotheses; (ii) a generalized hand-shadow alignment module to infer coarse hand poses with a similarity-driven strategy for selecting hypotheses; and (iii) a shadow-feature-aware refinement module to optimize the hand poses for physical plausibility and shadow feature preservation. Further, we design our pipeline to be trainable on generic public hand data, thus avoiding the need for any specialized training dataset. For method validation, we build a benchmark of 210 diverse shadow shapes of varying complexity and a comprehensive set of metrics, including a novel DINOv2-based evaluation metric. Through extensive comparisons with multiple baselines and user studies, our approach is demonstrated to effectively generate bimanual hand poses for a large variety of hand shapes for over 85% of the benchmark cases.

CCS Concepts: • **Applied computing** → **Media arts**; • **Computing methodologies** → *Shape modeling*.

Additional Key Words and Phrases: Shadow art, 3D hand pose estimation, visual art, computational art design, learning, generative posing

ACM Reference Format:
Hao Xu, Yinqiao Wang, Niloy J. Mitra, Shuaicheng Liu, Pheng-Ann Heng, and Chi-Wing Fu. 2025. Hand-Shadow Poser. *ACM Trans. Graph.* 44, 4 (August 2025), 16 pages. https://doi.org/10.1145/3730836

1 INTRODUCTION

Hand shadow art, also known as shadowgraphy [Nikola 1913], is a captivating art form, in which the shadows cast by hands on a wall creatively reveal the shapes of various kinds of objects. This art has a long and rich history across many cultures, since ancient times [Almoznino 1970; Jacobs 1996]. Its appeal lies in its simplicity, flexibility, and creativity. With only a few easy-to-obtain items (*i.e.*, hands, a light source, and a projective surface), one can create a

*indicates joint first authors.

Authors' addresses: Hao Xu; Yinqiao Wang, The Chinese University of Hong Kong, Hong Kong, China, {xuhao,yqwang}@cse.cuhk.edu.hk; Niloy J. Mitra, University College London, Adobe Research, London, United Kingdom, n.mitra@cs.ucl.ac.uk; Shuaicheng Liu, University of Electronic Science and Technology of China, Chengdu, China, liushuaicheng@uestc.edu.cn; Pheng-Ann Heng; Chi-Wing Fu, The Chinese University of Hong Kong, Hong Kong, China, {pheng,cwfu}@cse.cuhk.edu.hk.

0730-0301/2025/8-ART
https://doi.org/10.1145/3730836

(a) Western book (b) Chinese show (c) Japanese theatre group

Fig. 2. Hand-shadow examples from (a) the book "The art of hand shadows" [Almoznino 1970], (b) traditional shadow play [Shen 2024], and (c) theatre group [Gekidan Kakashiza 1952], from both the east and west.

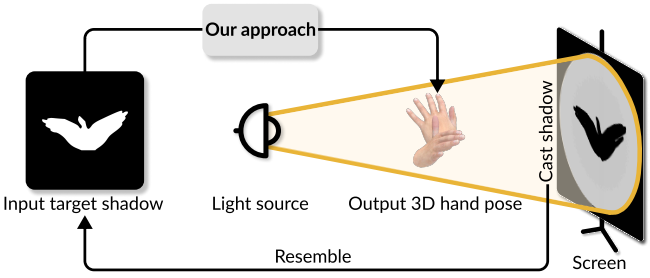


Fig. 3. Illustrating our task. Given a target shadow as the input, we aim to estimate the 3D poses of both the left and right hands, such that the two hands together can cast a shadow that closely resembles the input. Note that the light source and screen are fixed in the setup.

wide variety of shadow shapes that mimic lifelike animals, plants, portraits, *etc.*; see some classic examples in Figure 2.

We are interested in solving an inverse problem; see also Figure 3. We develop a learning-based approach to find plausible bimanual hand poses that can cast shadows that closely resemble a given hand shadow mask. Our approach enables users to explore a wide range of hand shadow forms, including animal-type shadows and extending to alphanumeric characters and even more intricate shapes, as shown in Figure 1 for some of our results.

Finding bimanual hand poses for reproducing a target hand shadow is nontrivial. First, the process is inherently ambiguous: the design space of 3D hand poses is huge, as a single shadow can often be produced by multiple different hand poses. Second, we need to attend to both the shape and crucial features in the input, but the absence of color and texture in shadows makes the hand shape recovery ill-posed (*i.e.*, significant changes in hand poses may not lead to any shadow changes, resulting in large plateau regions during pose-optimization via differentiable rendering). The fine-grained preservation of shadow features with restricted hand anatomy poses an additional challenge. Moreover, from a model learning perspective, the scarcity of domain-specific shadow datasets further complicates the method design. A detailed elaboration on the problem definition, setup, and challenges can be found in Section 3.

To meet these challenges, we build on one key insight: The inverse hand shadow art problem can be addressed through two subtasks. Given a hand shadow mask, by (i) resolving the *anatomical constraints* of two hands, it becomes feasible to recover anatomically correct hand shapes and poses; and (ii) resolving the *semantic constraints* of shadow allows the coarse hand poses to reproduce a shadow shape that preserves the features of the input. The decoupling allows solving the problem using only generic data with admissible hand configurations.

Specifically, to locate two hands from a bimanual hand mask, we first need to identify plausible 2D shapes for the left and right hands. It is challenging due to unknown overlapping regions and the mirror symmetry of the left and right hands. Deterministic methods like segmentation are suboptimal, as they cannot account for diverse possible hand configurations. We address this with a probabilistic generative model to produce diverse but reasonable hand assignments. Second, although shadows lack colors and textures, they provide shape priors. To generalize single-hand pose recovery to the shadow-mask domain, we fine-tune an RGB-based hand pose recovery model in a semi-supervised manner, leveraging existing knowledge while addressing the absence of 3D annotations. Last, to ensure that the reconstructed poses respect the most salient shadow features while maintaining anatomical plausibility, the optimization should prioritize key areas over the pixel-perfect alignment.

Based on these technical motivations, we design Hand-Shadow Poser, a three-stage framework to decouple the hand semantics from the anatomical constraints (imposed by the hands) and semantic constraints (imposed by the shadow shape): (i) A generative hand assignment module to predict plausible left-right hand shapes from the ambiguous shadow, by exploring diverse hypotheses via a conditional generative model. (ii) A generalized hand-shadow alignment module to robustly recover 3D poses of each hand-shape hypothesis to coarsely align with the shadow, followed by a similarity-driven strategy for selecting high-quality candidates. (iii) A shadow-feature-aware refinement module to iteratively optimize hand poses to reproduce salient features of shadow shape and ensure physical feasibility through carefully-designed constraints. Our feed-forward models in the first two stages are trained on generic public hand datasets with a rich set of augmentation operations, freeing us from creating extensive specialized hand-shadow data for training.

To evaluate our approach, we built a benchmark, containing diverse 2D masks of varying complexity, including shadow arts from books, alphanumeric characters, and everyday objects from the MPEG-7 dataset [Sikora 2001]. We also define a comprehensive set of metrics to assess the quality of the reproduced shadows, focusing on perception, semantics, and salient characteristics. Quantitative comparisons with baselines, qualitative results, and user studies consistently exhibit the effectiveness and robustness of our approach.

Overall, our main contributions are as follows:

- We introduce a comprehensive framework to compute hand shadow arts, covering a rich variety of shadow shapes.
- We design a three-stage pipeline to decouple the anatomical constraints imposed by the hand and the semantic constraints imposed by the shadow shape, enabling training merely on richly augmented generic public hand datasets.
- We formulate three novel components in our pipeline: generative hand assignment, generalized hand-shadow alignment, and shadow-feature-aware refinement.
- We construct a benchmark for evaluation, encompassing 210 shadow art forms with a variety of shapes, and introduce

shadow-specific metrics for quality assessment. The effectiveness of our approach is demonstrated through extensive experiments, including quantitative evaluations, qualitative comparisons, and detailed user studies.

The code and benchmark data of Hand-Shadow Poser will be publicly available at https://github.com/hxwork/HandShadowPoser.

2 RELATED WORK

Computational visual art. Visual arts embrace a wide variety of genres, media, and styles, demanding profound human aesthetics and expertise in creations [Wang et al. 2024]. A growing research has enabled computational generalization of various forms of visual arts, both 2D and 3D. For example, the generation of stylized artworks such as 2D paintings [Binninger and Sorkine-Hornung 2024; Chiu et al. 2015; Kopf and Lischinski 2011], 3D scenes [Haque et al. 2023; Liu et al. 2024; Zhang et al. 2022], 3D sculptures [Liu et al. 2017; Yang et al. 2021a], and reliefs [Schüller et al. 2014].

In shadowgraphy, shadows projected onto the wall present expressive shapes and figures (*e.g.*, animals), making it hard to believe that the shadow objects come merely from two human hands. This line of art differs from general visual arts. It is a specific genre characterized by a visual percept that differs from reality, such as optical illusion design [Coren 1978]. Specifically, our task lies at the intersection between 2D and 3D optical illusions.

2D optical illusion. Numerous computational methods have been proposed to synthesize illusional images. Oliva et al. [2006], in an early attempt, generate hybrid images that exhibit appearance changes at different viewing distances; Chi et al. [2008] propose to arrange repeated asymmetric patterns to stimulate illusory motion perception. Multiple efforts [Chu et al. 2010; Zhang et al. 2020; Zhao et al. 2024] study camouflage, in which the goal is to compute images with certain imagery patterns subtly embedded in the images. [Burgert et al. 2024; Geng et al. 2024] explore multi-view illusion images whose appearance changes upon flips, rotations, skews, or jigsaw rearrangements. Recently, Geng et al. [2025] generalize this technique to color saturation, motion blur, and inverse problems.

3D optical illusion. Computational generation of 3D optical illusions can be roughly divided into two categories. The first focuses on the digital fabrication of local microfacets for pattern display, *e.g.*, requiring certain lighting conditions. Various mediums have been exploited such as spatially-varying reflectance functions [Matusik et al. 2009], 3D height fields [Weyrich et al. 2009; Wu et al. 2022], microstructural stripe patterns [Sakurai et al. 2018], cellular mirrors [Hosseini et al. 2020], scratches on metal [Shen et al. 2023], and refractive lenses [Papas et al. 2012; Zeng et al. 2021]. Recently, researchers [Perroni-Scharf and Rusinkiewicz 2023; Zhu et al. 2024] exploit self-occlusion to achieve view-dependent appearances without relying on an external light source.

Our work is more related to the second category, which aims to generate a 3D shape that produces different forms of visual illusion. Gal et al. [2007] abstract input models into expressive 3D compound shapes with elements from a database. Leveraging depth misperception caused by projection, Wu et al. [2010] create topological structures that seem impossible to exist, whereas Sugihara [2014] creates solid shapes with slopes that appear to disobey the laws of gravity when a ball is placed on them. Tong et al. [2013] study the hollow-face illusion, in which a gradual deformation can be observed when walking around the object. Alexa and Matusik [2010] study reliefs that approximate given images under certain illumination; Chandra et al. [2022] design a differentiable probabilistic programming language to create multiple illusions, including human faces that appear to change expressions under different lighting. Creating 3D shapes with varying appearances from different view directions is initially explored in [Sela and Elber 2007], which relies on geometric deformation from two input 3D models. Keiren et al. [2009] provide a theoretical analysis of the problem of constructing a triplet from a given set of three letters. Intriguing variants are further studied, *e.g.*, in 3D shadow volumes [Mitra and Pauly 2009], 3D crystals [Hirayama et al. 2019], and 3D wire sculptures [Hsiao et al. 2018; Qu et al. 2024; Tojo et al. 2024].

Shadow art. Shadows are the results of the interplay between light and objects. Shadow art has been extensively explored to create expressive and illusional designs. Pellacini et al. [2002] present an interface for transforming shadows based on user requirements, whereas Mattausch et al. [2013] manipulate rendered shadows and apply the edited results in varying scene configurations.

Mitra et al. [2009] design an algorithm to construct a 3D volume constrained by orthogonal shadow images as inputs, such that lighting the same solid from different specific directions interestingly creates different shadow patterns. With a similar goal, Zhang et al. [2017a] develop a method to create 3D shadow art sculptures using a collection of real items. Chen et al. [2017] propose a framework for generating animated target shadows using objects under ballistic motion. Sadekar et al. [2022] revisit shadow art with a differentiable rendering-based optimization. Wang et al. [2024] further expand its potential and flexibility with implicit representations and joint optimization of lighting directions and screen orientations. A special form, namely creating 3D wire sculptures based on multi-view sketches, is first explored by Hsiao et al. [2018]. Recently, Qu et al. [2024] utilize flexible drawing capabilities from modern generative models, whereas Tojo et al. [2024] promote the fabricability of the reconstructed wires for 3D printing and support richer input controls. Gangopadhyay et al. [2024] deform a topological embedding of the circle in 3D space to create single- or multiple-view target shadows.

Instead of projecting from 3D shapes, some works create shadow art with manufactured planar-like devices. Alexa and Matusik [2012] design a planar surface with holes to create self-shadows that induce single input images. Bermano et al. [2012] exploit walls and chamfers within a diffuse surface for producing self-shadowing effects that display multiple images under different views and lights. Some variants study the casting of a single shadow onto an external plane to match various desired images. Baran et al. [2012] present a multilayer attenuator that casts different shadows depending on the light configuration. To project and form a target pixel art image, Yue et al. [2012] arrange transparent sticks within a container to refract light; Zhao et al. [2016] design 3D-printed perforated lampshades to project continuous grayscale images, whereas Min et al. [2017] arrange multiple occluder layers to create a soft boundary shadow.

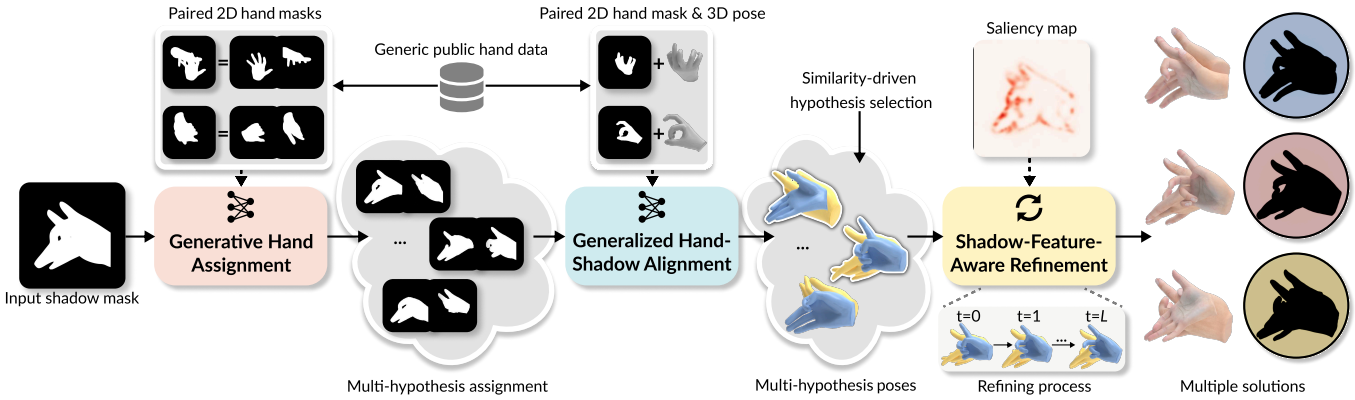


Fig. 4. Overview of our Hand-Shadow Poser, which consists of three key stages: (i) generative hand assignment, (ii) generalized hand-shadow alignment with similarity-driven hypothesis selection, and (iii) shadow-feature-aware refinement.

In this work, we aim to create prescribed shadows using human hands, inspired by the traditional hand shadow arts [Almoznino 1970; Jacobs 1996; Nikola 1913]. One of the most relevant works is [Won and Lee 2016], which generates human characters given 2D silhouette images by employing a nonlinear optimization to minimize the visual difference between the resultant and target shadow contours. However, they require hints specified by professional actors to match specific points on the target contour with associated body parts, which is difficult to obtain for various shadow inputs in real-world scenarios. Besides, directly applying a method following [Won and Lee 2016] in our task tends to yield unsatisfactory results, since the optimization is inherently sensitive to initialization and easily converges to a local optimum.

Another closely related work is a short paper [Gangopadhyay et al. 2023], which solves a similar task to ours. They use differentiable rendering and directly minimize the image loss between the input and target shape. Result-wise, it showcases only a few hand shadow examples. Due to its optimization-based nature, similar issues are observed as in [Won and Lee 2016]. Beyond the above two works, we present a novel and generalizable approach capable of (i) covering a richer variety of hand shadow cases, (ii) capturing salient characteristics of the target shadow, and (iii) generatively proposing diverse hand poses with anatomical constraints. Also, we take optimization through differentiable rendering as a baseline in our comparison and show that differentiable rendering alone cannot achieve the results of our approach; see Section 7 for the comparison experiment. To our best knowledge, this is the first work that comprehensively studies the creation of hand shadow arts.

3D hand pose estimation from silhouettes. Another closely-related research topic is 3D hand pose estimation from monocular RGB images [Chen et al. 2022; Huang et al. 2023; Iqbal et al. 2018; Moon and Lee 2020; Pavlakos et al. 2024; Xu et al. 2023; Zhang et al. 2021, 2019; Zhou et al. 2020, 2024], a longstanding research task due to its significance in downstream applications. Yet, very few attempts have been made to recover 3D hand poses from sparse 2D information, such as anatomical landmarks [Ramakrishna et al. 2012], hand-drawn stick figures [Lin et al. 2012], or binary masks [Agarwal and Triggs 2004; Dibra et al. 2017] that are conducted on human bodies. [Lee et al. 2019] is the first work that estimates 3D single-hand pose from binary silhouettes, which requires additional depth supervision during the training stage. Under the same setting, Chang et al. [2023] achieve comparable performance as state-of-the-art RGB-based and depth-based methods without relying on depth information. However, both works focus on single-hand inputs. Directly applying their method to bimanual hand masks remains challenging since we need to estimate the locations of the two hands in the input while the input is simply a binary mask, in which the hand shapes are obscured. Thus, we should not only solve the ill-posed problem of locating a pair of non-intersecting interacting hands within a single mask, but also collectively estimate the poses of the two hands to reproduce the target shadow.

3 OVERVIEW

Problem definition. Figure 3 illustrates our task. The input is a target shadow represented as a binary mask, whereas the outputs are the 3D poses of the left and right hands represented by the MANO [Romero et al. 2022] hand model. With a light source and screen plane, we aim to inversely find the 3D poses of the hands positioned between them, such that the projected hand shadow on the screen can closely match the given target shadow.

We further clarify the setup. Creating hand shadows with clear and sharp boundaries requires a small, intense light source and a flat projection screen, as outlined in classical references [Almoznino 1970; Nikola 1913]. The light source and screen remain fixed, while the hands are adjusted in between. The hands, light source, and screen are horizontally and vertically aligned to minimize distortion. For simplicity, we focus on scenarios with only two hands, without considering other body parts and additional object items.

Challenges. To achieve our goal, one straightforward approach is to directly optimize the hand poses by minimizing the visual difference between the cast shadow and the target shadow [Gangopadhyay et al. 2023; Won and Lee 2016]. However, there exist several key challenges outlined below:

(i) *Initialization sensitivity*: Optimization-based methods are sensitive to initialization and prone to converge to local optima. Providing a good initial condition is a crucial step towards a

successful result and fast optimization [Finn et al. 2017], yet it remains challenging due to the huge search space.

(ii) *Feature preservation*: It is infeasible to match every pixel of the projected and the input shadows due to limited hand anatomy. Instead, the most distinctive features of the input mask should be retained, whereas manually specifying hints [Won and Lee 2016] is impractical.

Another straightforward approach is to train a neural network model to predict the interacting hand poses from the target shadow in a feed-forward manner, utilizing prior distributions learned from training datasets. Yet, this process is also nontrivial due to the following challenges:

(iii) *Dataset scarcity*: Given the scarcity of annotated hand shadow art datasets, the model must be robust and generalizable, without relying on prior knowledge from a specific data domain, to avoid labor-intensive data preparation.

(iv) *Results diversity and robustness*: The same given shadow could be produced by multiple different hand poses. Especially when two (left and right) hands are considered, there can be many different choices. Identifying diverse yet reasonable results introduces another challenge to the network design.

Overview of our Hand-Shadow Poser. Figure 4 gives an overview of our approach, which has the following three stages: (i) the *generative hand assignment* stage assigns diverse reasonable left-right 2D hand shapes (masks) to cover different parts of the shadow in the input binary mask (Section 4); (ii) the *generalized hand-shadow alignment* stage recovers a coarse 3D hand pose of each single-hand binary mask and automatically selects the high-quality ones for the subsequent stage (Section 5); and (iii) the *shadow-feature-aware refinement* stage iteratively refines the coarse 3D hand poses to make their shadows resemble the input, considering physical plausibility (Section 6). In the end, we take our approach to work on diverse hand shadow examples from our benchmark, and conduct a series of evaluations to demonstrate the quality of our results and the effectiveness of the proposed designs.

4 GENERATIVE HAND ASSIGNMENT

The first stage aims to find rough 2D hand shapes (masks) with reasonable anatomy to match the input shadow. We name this task *hand assignment*, *i.e.*, to assign each hand to cover different parts of the target shadow. In particular, we do not require the 3D hand poses for shadow matching in this stage. Here, the main challenges are due to the lack of information in the input, which is just a binary mask, and also to the many different possible hand shapes that may eventually match and form the target shadow.

An initial attempt. At the beginning of this research, we tried an image segmentation approach, *i.e.*, to classify each pixel in the input shadow mask as the left hand, right hand, or both (overlapping). Specifically, we adopted the network architecture in [Liu et al. 2023] and trained it on a mixture of datasets with rendered segmentation labels. Then, we observed several drawbacks. First, due to the deterministic nature, using a segmentation model discourages capturing the uncertainty in shadow-to-hand mapping. Second, segmentation emphasizes pixel-level accuracy, so the trained model tends to focus excessively on the hand shapes rather than exploring their global cues. Last, the absence of color and texture in the input largely raises the complexity of network learning compared to conventional segmentation tasks. Hence, this approach leads to inferior performance, as shown later in Section 7.

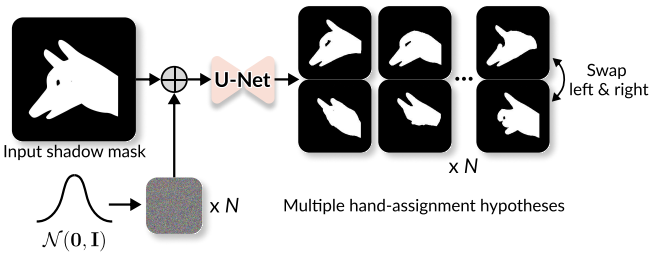


Fig. 5. Our generative approach for hand assignment introduces diversity for exploring more different 2D hand shapes. Likewise, we additionally swap the left and right hands to model mirror symmetry in hand assignment.

Our generative approach. To overcome the above issues, we propose to formulate a generative approach for hand assignment. By doing so, we aim to introduce diversity in the results to address the ambiguity in shadow-to-hand mapping; see Figure 5. Also, we aim to make the network learning easy, so that the network model can better attend to the overall hand shapes than pixel-level recovery.

Method-wise, we design the generative hand assignment model based on the conditional denoising diffusion probabilistic model [Ho et al. 2020]. To learn the reverse diffusion process, we adopt the classifier-free guidance [Ho and Salimans 2022] for shadow-controlled multi-hypothesis generation. By concatenating the input binary mask $\hat{\mathbf{M}}$ with the intermediate noisy output $\mathbf{x}_t$ at timestamp t (ranging from 0 to T), our network model can progressively reach the final assignment $\mathbf{x}_0 \in \mathbb{R}^{H\times W\times 2}$ (*i.e.*, a two-channel image with height H and width W) using the denoising model $f_{\text{assign}}(\cdot)$:

$$\mathbf{x}_0 = f_{\text{assign}}(\hat{\mathbf{M}}, \text{PE}(t)) \tag{1}$$

where t is encoded through positional embedding (PE) [Vaswani et al. 2017]. The assigned left- and right-hand masks $\mathbf{M}_l$ and $\mathbf{M}_r$ are then obtained from $\mathbf{x}_0$ using

$$\mathbf{M}_l, \mathbf{M}_r = \text{Split}(\mathbf{x}_0), \tag{2}$$

where $\text{Split}(\cdot)$ denotes the channel split operation.

In addition, to speed up the inference, we employ DDIM [Song et al. 2020] to sample $\mathbf{x}_t$ at arbitrary timestamps. At inference, N initial noise vectors $\{\mathbf{x}_T^i | i \in 1, ..., N\}$ are randomly sampled from the Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ to produce diverse hand assignment results. Also, we swap the left and right hands (Figure 5) to further enrich the diversity. Specifically, a U-Net is adopted as the denoising model $f_{\text{assign}}(\cdot)$, in which we employ an encoder with four downsample blocks, each with two residual blocks; an attention mechanism; a downsampling layer; and a decoder with four upsample blocks in a structure similar to the encoder. Further, a residual block is employed to yield the two-channel map $\mathbf{x}_0$.

Furthermore, following [Karras et al. 2022], we employ the L2 distance between the predicted and ground-truth values as the training

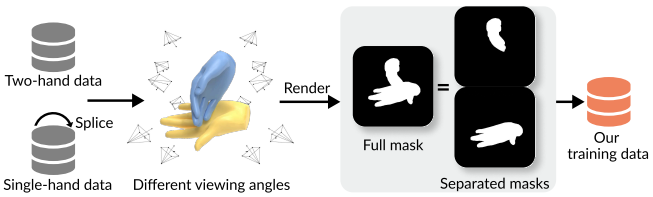


Fig. 6. To prepare training data for generative hand assignment, we augment existing two-hand and single-hand datasets by (i) randomly splicing left- and right-hand samples in single-hand datasets to synthesize more two-hand samples, and (ii) rendering two-hand samples in different views.

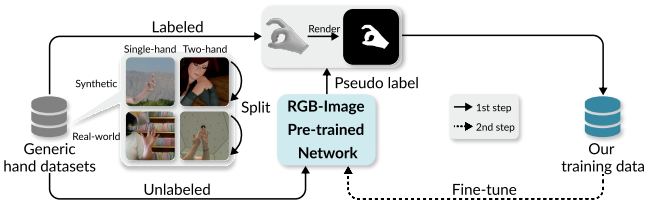


Fig. 7. To prepare training data for generalized hand-shadow alignment, we propose to use a semi-supervised learning strategy as illustrated above, considering the use of both labeled and unlabeled hand samples from both real and synthetic datasets. With this strategy, we can produce a dataset with paired 2D masks and 3D poses for fine-tuning our network.

loss, in which we separately calculate the left and right masks:

$$\mathcal{L}_{\text{assign}} = \frac{\bar{\alpha}_t}{1-\bar{\alpha}_t} \sum_{*\in\{l,r\}} ||\mathbf{M}_* - \hat{\mathbf{M}}_*||_2^2, \tag{3}$$

where $\bar{\alpha}_t$ denotes the total noise variance at step t, as defined in [Song et al. 2020].

Training data. Benefiting from our decoupled pipeline design, the hand assignment model mainly needs to learn the knowledge of 2D hand shapes instead of shadow semantics. Hence, to prepare for the training data, we propose to leverage the rich 2D and 3D ground-truth labels in existing public hand datasets [Moon et al. 2020; Zhang et al. 2017b; Zimmermann et al. 2019; Zuo et al. 2023] that were built for various other purposes *e.g.*, hand pose estimation and tracking. By doing so, we can train our model using generic hand datasets, including also synthetic ones [Li et al. 2023].

To do so, we augment existing hand datasets to provide the supervision for network model training in two aspects, as illustrated in Figure 6. First, since two-hand datasets are scarce, compared with single-hand ones, we randomly splice (combine) left- and right-hand samples from single-hand datasets, following [Zuo et al. 2023], thereby synthesizing more diverse interacting poses of varying levels of hand overlap, which could occur in real hand shadow art scenarios. Second, we render the 3D interacting hand meshes from multiple perspectives to enrich the diversity of viewing angles. By these means, we can substantially increase both the quantity and diversity of two-hand samples for network training.

5 GENERALIZED HAND-SHADOW ALIGNMENT

Given the estimated left- and right-hand masks $\mathbf{M}_l$ and $\mathbf{M}_r$, the second stage aims to construct the 3D poses (*i.e.*, the 61 MANO coefficients that represent the hand orientation, axis-angle 3D poses of 15 hand joints, hand shape, and 3D coordinate of the wrist joint) of the left and right hands (θ_l, β_l, t_l) and (θ_r, β_r, t_r), such that the resulting hand poses provide a coarse 3D hand alignment with the target shadow. Importantly, we do not require capturing the fine-grained shadow features at this stage. Rather, we need coarse 3D predictions from rough 2D hand shapes of diverse poses.

Considering that single-hand poses are relatively easier to infer than collectively estimating interacting hand poses, we propose to narrow down the search space by predicting the pose of each hand separately. Here, we train the pose recovery network $f_{\text{align}}(\cdot)$ to predict the MANO representation of each hand from its mask:

$$\theta_*, \beta_*, t_* = f_{\text{align}}(\mathbf{M}_*), \tag{4}$$

where subscript $*$ denotes l (left) or r (right). Yet, to make a good prediction is still nontrivial for two reasons. First, it is hard to recover 3D hand poses from the colorless masks, providing only sparse information. Second, the input hand mask is simply a rough approximation of the actual hand shape; due to shadow ambiguity, we need a robust network model to overcome the uncertainty.

A generalized approach. To meet these challenges, we aim for a generalizable and robust performance from two perspectives: (i) generalizing well-learned knowledge from the RGB-image domain to the shadow mask domain; and (ii) leveraging large data priors in existing data to generalize and handle rough 2D hand shapes.

The above considerations motivate us to adopt a large-scale fully transformer-based design [Dosovitskiy 2020]. First, to fully generalize knowledge from the RGB image domain, we initialize the network model f_{align} with the pre-trained weights from [Pavlakos et al. 2024] to leverage vast data prior learned from extensive RGB image data. To effectively handle the uncertainty in the inputs, we further fine-tune the network using a comparable magnitude of binary mask images from a large collection of generic hand datasets, including both real and synthetic data with single and interacting hands. Specifically, we adopt the Vision Transformer (ViT) [Dosovitskiy 2020] as the network backbone, which takes embeddings of image patches as input. The output tokens are then fed into a transformer decoder to regress the MANO parameters by cross-attending to a single query token. Last, the hand mesh and its relative translation to the camera can be converted through a MANO layer.

Model training. We adopt loss functions similar to [Dosovitskiy 2020] to supervise the network training:

$$\begin{aligned} \mathcal{L}_{\text{align}}^{\text{3D}} &= \|\theta - \hat{\theta}\|_2^2 + \|\beta - \hat{\beta}\|_2^2 + \|\mathbf{J}^{\text{3D}} - \hat{\mathbf{J}}^{\text{3D}}\|_1 \\ \text{and } \mathcal{L}_{\text{align}}^{\text{2D}} &= \|\mathbf{J}^{\text{2D}} - \hat{\mathbf{J}}^{\text{2D}}\|_1, \end{aligned} \tag{5}$$

where $\mathbf{J}^{\text{3D}}$ denotes the 3D joint coordinates converted from the predicted MANO parameters; $\mathbf{J}^{\text{2D}}$ denotes their projections onto the image space by using the camera intrinsics; and the quantities with the hat superscript $\hat{}$ are the ground-truth labels. Here, $\mathcal{L}_{\text{align}}^{\text{2D}}$ is utilized to promote consistency in the output image space, following [Dosovitskiy 2020].

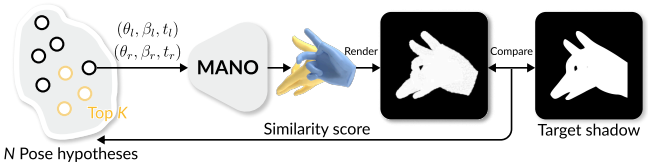


Fig. 8. Our similarity-driven hypothesis selection strategy, taking a render-and-compare approach to favor pose hypotheses of the highest quality.

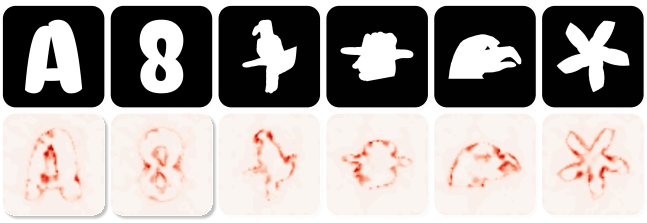


Fig. 9. Top row: target shadows. Bottom row: extracted saliency maps, in which characteristic shadow features are highlighted in red.

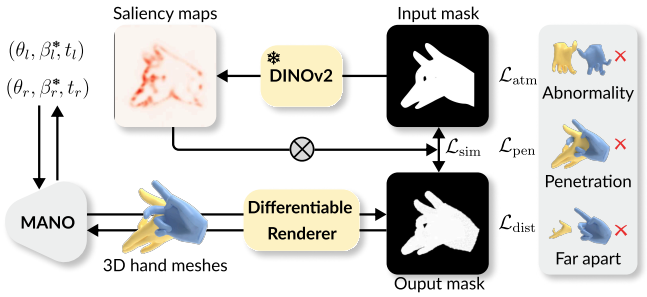


Fig. 10. Our shadow-feature-aware refinement iteratively optimizes the 3D hand poses to align with features in the input using the DINOv2-based saliency guidance, while considering physical constraints in terms of anatomy, penetration, and hand-to-hand distance.

In the training process, we first split the left and right hands from the existing two-hand data to obtain more single-hand training samples. Since not all data samples are paired with MANO-represented ground truths, we take a semi-supervised learning strategy as illustrated in Figure 7, *i.e.*, employing the RGB-image pre-trained network on unlabeled images to estimate the MANO coefficients as pseudo labels, then rendering the results to produce paired binary hand masks. With this approach, we can avoid the need for highly accurate pseudo labels associated with the original images, as our focus is on ensuring the anatomical correctness of the hand poses. Second, we take these samples, together with the labeled samples, to form our dataset for network training.

Similarity-driven hypothesis selection. The 2D hand shape hypotheses from the previous stage are fed into the pose recovery network to obtain 3D hand poses, *i.e.*, pose hypotheses. However, this process does not take into account the quality of the hypotheses, which may largely degrade the performance of the next stage. Since ground truths are not available at inference, we thus formulate a similarity-driven strategy to evaluate and select pose hypotheses.

Overall, our idea is to maximize the similarity between the target shadow and the reproduced shadow, by a render-and-compare approach. That is, we first project and render each pose hypothesis (*i.e.*, its 3D hand mesh) into a binary mask, and then calculate its perceptual similarity to the input mask using LPIPS [Zhang et al. 2018] and DINOv2 [Oquab et al. 2023] semantic-based scores. By sorting all N hypotheses based on their similarity scores, we can then select the top K hypotheses for refinement in the next stage; see Figure 8. Details about similarity scores are introduced in Section 7.1.

6 SHADOW-FEATURE-AWARE REFINEMENT

To successfully reproduce the target shadow, the final 3D hand poses need to attend to the global shape of the shadow, as well as to the details or shadow features; see *e.g.*, the beak of the parrot and the eyes of the eagle and wolf in Figures 9 and 10. Hence, the final stage aims to refine the coarse 3D hand poses to make their projections perceptually more similar to the target shadow, considering particularly the shadow features together with the anatomical constraints.

Method-wise, the overall approach is based on differentiable rendering. That is, we first create a binary mask of the hands by projecting the coarse 3D hand meshes from the previous stage. Then, we iteratively optimize the joint angles and wrist positions of the two hands with their shape parameters fixed, mimicking real-world hand pose adjustments; see Figure 10. Importantly, beyond the differentiable rendering in [Gangopadhyay et al. 2023], where pose initialization is rarely considered and often leads to suboptimal results, the coarse outputs from our first two stages provide a good initial condition for the optimization process to achieve a faster and better convergence [Finn et al. 2017; Lee et al. 2020; Rajeswaran et al. 2019]. This can be attributed to the prior knowledge of hands brought about by our decoupling design.

Below, we introduce four carefully-crafted constraints for the optimization. The first constraint aims to maximize the similarity between the input and rendered masks, with saliency guidance for preserving the shadow features. To favor physically-plausible hand poses, we further incorporate the other three constraints, considering anatomy, penetration, and hand-to-hand distance.

(i) Similarity constraint with saliency guidance. In the optimization, our main goal is to minimize the misalignment between the rendered mask $\mathbf{M}$ and the input mask $\hat{\mathbf{M}}$. Directly constraining their image discrepancy with L1 or L2 loss can lead to suboptimal results, due to the significant gap in flexibility between the limited range of hand joint movement and the expressive capacity of shadows. Rather than aligning every pixel equally, we prioritize preserving the shadow features. Also, this process is desired to be automated, eliminating the need to manually specify the hint points in [Won and Lee 2016]. To this end, we propose to leverage DINOv2 [Oquab et al. 2023], a powerful pre-trained vision model, to first locate prominent features in the input shadow shape; see again Figure 9). Given the input mask $\hat{\mathbf{M}}$, we assign varying levels of importance to different shadow regions based on the extracted saliency map:

$$\mathcal{L}_{\text{sim}} = \sum \left(1 + \text{DINO}(\hat{\mathbf{M}})\right) \odot \left|\mathbf{M} - \hat{\mathbf{M}}\right|, \tag{6}$$

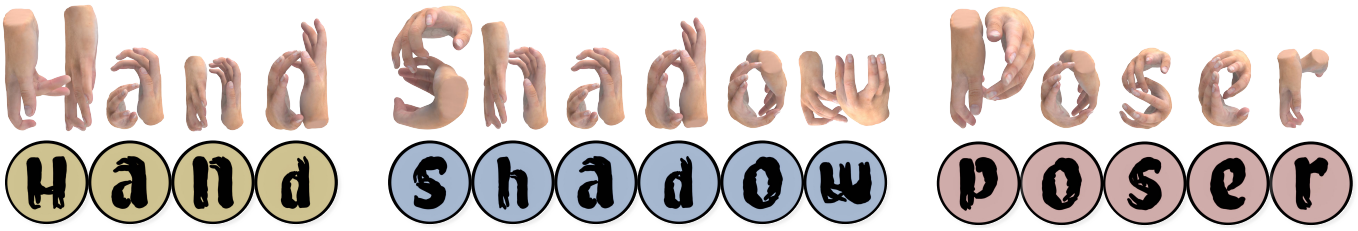


Fig. 11. A gallery of "Hand Shadow Poser" created by our Hand-Shadow Poser on various uppercase and lowercase letters.

where DINO$(\cdot)$ represents DINOv2's attention heatmap extraction and $\odot$ is the Hadamard product.

(ii) Anatomy constraint. To mitigate a pose's abnormality, we adopt the twist-splay-bend frame in [Yang et al. 2021b] by projecting the rotation axis to three independent axes, then computing the penalization of the abnormal axial components on each joint as $\mathcal{L}_{\text{atm}}$. Please refer to [Yang et al. 2021b] for the details.

(iii) Penetration constraint. Inspired by [Jiang et al. 2021], we identify vertices of one hand that are inside the other hand (the set is denoted as $\mathbf{P}_{\text{in}}$) and define the inter-penetration loss as their distances to the closest vertices on the other hand:

$$\mathcal{L}_{\text{inter-pen}} = \frac{1}{|\mathbf{P}_{\text{in}}|} \sum_{p \in \mathbf{P}_{\text{in}}} min_i \|p - \mathbf{V}_i\|_2^2, \tag{7}$$

where $\{\mathbf{V}_i\}$ represents mesh vertices of the hand being penetrated. For self-penetration, we adopt the conic distance fields approximation of meshes in [Tzionas et al. 2016] to penalize the depth of intrusion, denoted as $\mathcal{L}_{\text{self-pen}}$. The final penetration loss $\mathcal{L}_{\text{pen}}$ is a sum of $\mathcal{L}_{\text{inter-pen}}$ and $\mathcal{L}_{\text{self-pen}}$. For the detailed calculation of $\mathcal{L}_{\text{self-pen}}$, please refer to [Ballan et al. 2012; Tzionas et al. 2016].

(iv) Hand-to-hand distance constraint. With the above constraints, we optimize the validity of two hand poses and achieve the desired shape in the projection space. However, since the optimization is not sensitive to movements along the depth axis after the projection, the resulting hand meshes can become too far apart along the depth axis relative to the light source. Moreover, even a relatively moderate distance, *e.g.*, one meter, can significantly complicate the process of creating the hand shadows.

Concerning this, we propose a new loss term to constrain the distance between the wrist joints of the two hands. Empirically, we penalize this distance when it exceeds a certain threshold τ_{dist}.

$$\mathcal{L}_{\text{dist}} = \begin{cases} \|t_l - t_r\|_2^2 & \text{if } \|t_l - t_r\|_2^2 \geq \tau_{\text{dist}} \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

where t_l and t_r are the 3D joint coordinates of the left-hand and right-hand wrists, respectively.

Optimization. The final objective is a weighted sum of the terms:

$$\min_{\theta_l, t_l, \theta_r, t_r} [w_{\text{sim}}\mathcal{L}_{\text{sim}} + w_{\text{atm}}\mathcal{L}_{\text{atm}} + w_{\text{pen}}\mathcal{L}_{\text{pen}} + w_{\text{dist}}\mathcal{L}_{\text{dist}}]. \tag{9}$$

where w_{sim}, w_{atm}, w_{pen}, and w_{dist} are hyperparameters. Further, we adopt Adam [Kingma and Ba 2015] for gradient-descent-based optimization, which ends after L iterations.

7 RESULTS AND EXPERIMENTS

7.1 Experimental Setup

Baselines. We compare our Hand-Shadow Poser with three baselines:

- *Baseline 1* optimizes the 3D hand poses with differentiable rendering as in [Gangopadhyay et al. 2023], with random initialization three times, and then picks the best one based on the similarity metrics.
- *Baseline 2* uses a single neural network to directly regress the coarse pose of the interacting hands from the input shadow mask, followed by the same optimization as *Baseline 1*.
- *Baseline 3* replaces the generative model in Stage 1 with a segmentation model, as described in Section 4.

Training datasets. We prepare the training data for the feed-forward models from multiple public hand datasets, including single- and two-hand datasets. Specifically, to train the generative hand assignment model in Stage 1 (Section 4), we prepare pairs of two-hand masks and left-right hand masks from InterHand2.6M [Moon et al. 2020], RenderIH [Li et al. 2023], and Two-hand 500K [Zuo et al. 2023]. We also follow [Zuo et al. 2023] to randomly combine single-hand data in [Gomez-Donoso et al. 2019; Moon et al. 2020; Zhang et al. 2017b; Zimmermann and Brox 2017; Zimmermann et al. 2019]. Leveraging the multi-perspective augmentation strategy, we obtain 7.7M data samples in total for generative model training.

To train the large-scale transformer network in Stage 2 (Section 5), we use a large collection of public datasets, following [Pavlakos et al. 2024], FreiHAND [Zimmermann et al. 2019], HO3D [Hampali et al. 2020], MTC [Xiang et al. 2019], RHD [Zimmermann and Brox 2017], InterHand2.6M [Moon et al. 2020], H2O3D [Hampali et al. 2020], DexYCB [Chao et al. 2021], COCO WholeBody [Jin et al. 2020], Halpe [Fang et al. 2022], and MPII NZSL [Simon et al. 2017]. We additionally incorporate RenderIH [Li et al. 2023] and Two-hand 500k [Zuo et al. 2023] by splitting them into left- and right-hand data. The final training set consists of 2.7M samples.

Evaluation benchmark. We constructed a benchmark of 210 binary mask images, covering a wide variety of hand shadow shapes, for quantitative and qualitative evaluation. The dataset includes 62 alphanumeric characters (C1), 87 real hand-shadow-art shapes (C2) from [Almoznino 1970; Jacobs 1996; Nikola 1913], and 61 shapes of diverse everyday objects (C3) from [Sikora 2001] and Internet. For more examples, please refer to our supplementary material. To the best of our knowledge, this is the first work that collects such a

Fig. 12. A gallery showcasing the results of our Hand-Shadow Poser on real hand-shadow-art shapes (C2), which are obtained from the following books [Almoznino 1970; Jacobs 1996; Nikola 1913], covering a wide range of shapes encompassing animals, human portraits, buildings, and plants. For each case: the top left shows the target shadow, the bottom left shows our reproduced shadow, whereas the right shows our produced 3D hand poses.

diverse and challenging set of hand shadow shapes for a systematic analysis of computing hand-shadow arts.

Metrics. On the other hand, we propose to use the following five metrics to evaluate the visual similarity between the generated shadow and the input shadow: (i) *LPIPS*: We adopt LPIPS [Zhang et al. 2018] to measure the perceptual similarity based on the deep features from AlexNet [Krizhevsky et al. 2012]. Building on CLIP [Radford et al. 2021], we employ two other similarity metrics, including (ii) *CLIP-Global*, which evaluates the image-level semantic similarity by employing the CLIP image encoder to map the shadow mask image to the CLIP space and then calculating the cosine distance; and (iii) *CLIP-Semantic*, which computes the cosine similarity between the CLIP text embedding of the input shadow's class description (*e.g.*, "rabbit") and the image embedding of the generated shadow, to assess its level of alignment with the text semantics. Considering the model's sensitivity to text, we adopt the officially-released CLIP code by scaling the original similarities by a factor of 100, followed by a softmax operation to obtain the logit scores for the reproduced hand shadow mask of each baseline and our method, respectively. (iv) *DINO-Global*: Similar to *CLIP-Global*, we leverage DINOv2 [Oquab et al. 2023] for feature extraction to evaluate the visual similarity at a global scale. (v) *DINO-Semantic*: Additionally, to remedy the ignorance of the above metrics to local characteristics, we design this metric to measure the preservation of local shadow features between the output mask $\mathbf{M}$ and input mask $\hat{\mathbf{M}}$:

$$\textit{DINO-Semantic} = \frac{\sum(\mathbf{M} - \hat{\mathbf{M}}) \odot \mathbb{1}(\text{DINO}(\hat{\mathbf{M}}) > \tau_{\text{semantic}})}{\sum \mathbb{1}(\text{DINO}(\hat{\mathbf{M}}) > \tau_{\text{semantic}})}, \quad (10)$$

where τ_{semantic} is set to 0.1 by default and $\mathbb{1}(\cdot)$ is the indicator function. These metrics together provide a comprehensive evaluation of the quality of the reproduced shadows.

Implementation details. We adopt Blender [Blender 2019] to create all the hand-shadow-art scenes. For shadow projection, we set up a spotlight with a beam radius of 0.001 and an angle of 15°, with 1000 W power. The distance from the light source to the projection plane is set to 2.5 m. The focal length of the perspective camera in differentiable rendering is set to 1 m, aligning with the one used in [Pavlakos et al. 2024]. To avoid projection deviation, the camera is positioned at the same world coordinates as the light source, facing the same direction towards the projective surface. We implemented our method using PyTorch [Paszke et al. 2019] and adopted the Adam optimizer for both training the feed-forward models (Stages 1 and 2) and optimizing the hand orientations, poses, and translations (Stage 3). All experiments were conducted on eight NVIDIA Tesla V100 GPUs.

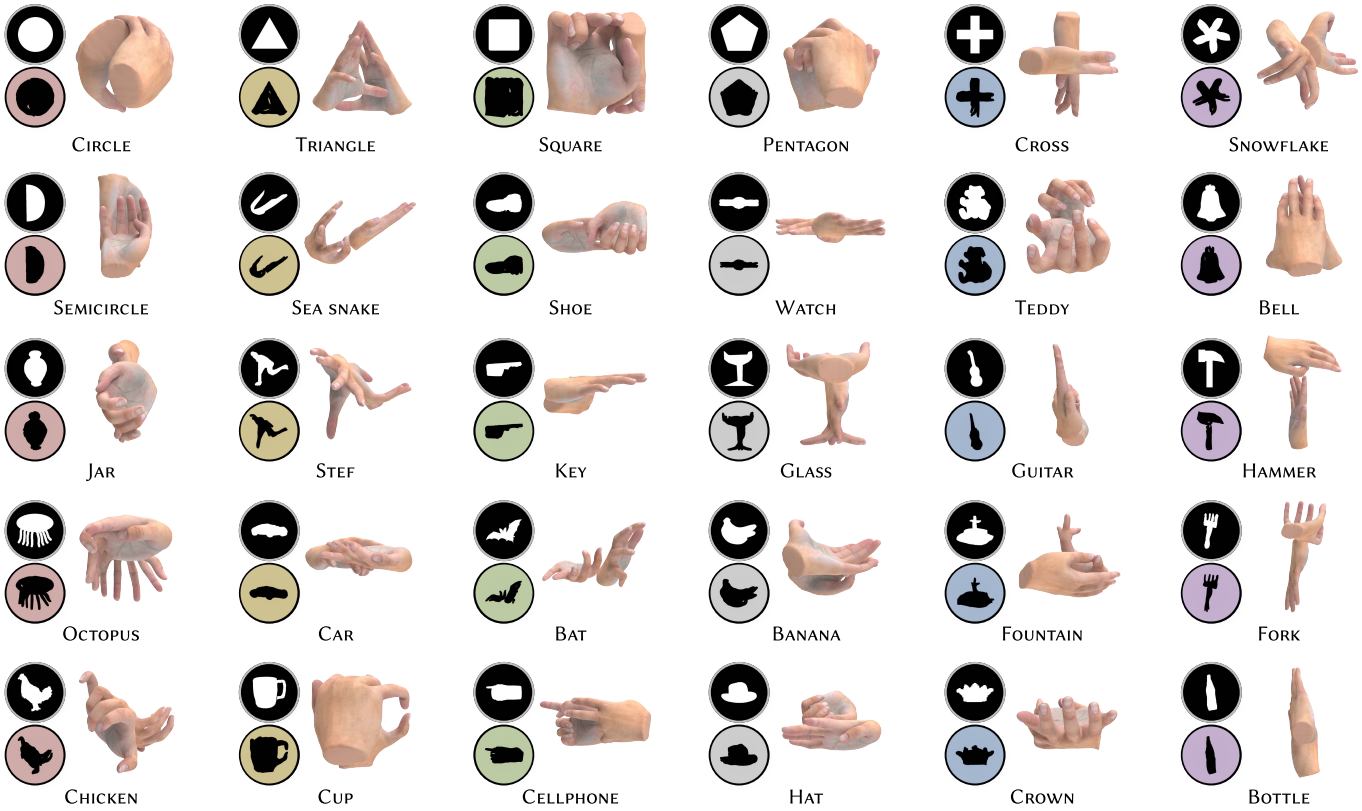


Fig. 13. A gallery of hand shadow arts created by our Hand-Shadow Poser for shapes of diverse everyday objects (C3) from [Sikora 2001] and the Internet. For each case, the top left shows the target shadow, the bottom left shows our reproduced shadow, whereas the right shows our produced 3D hand poses.

Fig. 14. Left: our physical setup. Right: real shadows created for Kangaroo, Crab, Deer, Camel, and Fox chases rabbit.

Fig. 15. 3D-printing seven of our 3D hand pose results, which are reconstructed for reproducing the following shadow shapes: Tortoise, Dinosaur, Farmer, Deer, Dolphin, Elephant, and Cat (left to right).

Specifically, for Stage 1, we use a batch size of 48 with a learning rate of 1e-4 and train the network model for 20 epochs. The input images are resized to 256×256, with a random rotation in $[0, 360^\circ]$ and scaling in $[0.75, 1.25]$ for online data augmentation. During the inference, the number of reverse steps for DDIM is 1,000.

In Stage 2, the model is fine-tuned for 10 epochs using a batch size of 8 and a learning rate of 1e-5. The input images are resized to 256×256 with the online augmentation strategies in [Pavlakos et al. 2024]. For similarity-driven hypotheses selection, we set the default number of candidate hypotheses N to 20 and the number of selected poses K to 3. The training processes for the first two stages take 3 days and 1 day, respectively.

For Stage 3, w_{sim}, w_{atm}, w_{pen}, and w_{dist} are empirically set to 10.0, 1.0, 1.0, and 1.0, respectively, whereas τ_{dist} is set to 0.5 by default. A Gaussian blur with a kernel size of 15×15 is applied to the extracted saliency map. We optimize the hand parameters with a learning rate of 1e-3 and decay it by 0.5 at the 3,000th iteration, with the maximum number of iterations L set to 6,000.

Table 1. Quantitative comparison between baselines and our Hand-Shadow Poser. C1: Alphanumeric characters; C2: Real hand-shadow-art shapes; and C3: Shapes of everyday objects. The first five columns present metrics (*LPIPs*, *CLIP-Global*, *CLIP-Semantic*, *DINO-Global*, and *DINO-Semantic*) used to evaluate the visual similarity, while the last two columns (*Human-Global* and *Human-Semantic*) are human-related metrics from the user study.

Methods	*LPIPS* ↓		*CLIP-Global* ↑		*CLIP-Semantic* ↑		*DINO-Global* ↑		*DINO-Semantic* ↓		*Human-Global* ↑		*Human-Semantic* ↑	
	C1 / C2 / C3	Avg.	C1 / C2 / C3	Avg.	C1 / C2 / C3	Avg.	C1 / C2 / C3	Avg.	C1 / C2 / C3	Avg.	C1 / C2 / C3	Avg.	C1 / C2 / C3	Avg.
Baseline 1	0.19 / 0.20 / 0.17	0.19	0.84 / 0.91 / 0.88	0.88	0.11 / 0.29 / 0.15	0.20	0.51 / 0.61 / 0.51	0.55	0.66 / 0.81 / 0.64	0.72	2.27 / 2.11 / 2.42	2.27	2.35 / 2.34 / 2.55	2.41
Baseline 2	0.20 / 0.19 / 0.17	0.19	0.83 / 0.91 / 0.88	0.88	0.10 / 0.18 / 0.22	0.17	0.47 / 0.63 / 0.51	0.55	0.70 / 0.84 / 0.66	0.74	2.03 / 2.49 / 2.36	2.29	2.15 / 2.42 / 2.42	2.33
Baseline 3	0.18 / 0.17 / 0.15	0.16	0.89 / 0.93 / 0.91	0.91	0.36 / 0.20 / 0.27	0.27	0.65 / 0.74 / 0.65	0.69	0.61 / 0.75 / 0.59	0.66	3.66 / 3.07 / 3.53	3.42	3.49 / 3.07 / 3.45	3.33
Ours	**0.15** / **0.15** / **0.13**	**0.14**	**0.91** / **0.95** / **0.93**	**0.93**	**0.42** / **0.33** / **0.35**	**0.36**	**0.71** / **0.80** / **0.75**	**0.78**	**0.47** / **0.67** / **0.49**	**0.56**	**4.66** / **4.45** / **4.47**	**4.53**	**4.55** / **4.21** / **4.29**	**4.35**

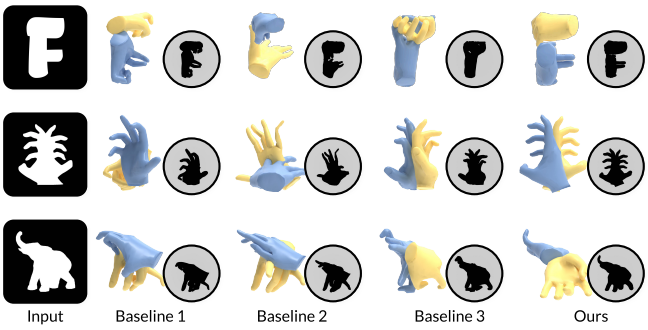


Fig. 16. Comparing hand-shadow-art results produced by the three baselines and by our Hand-Shadow Poser.

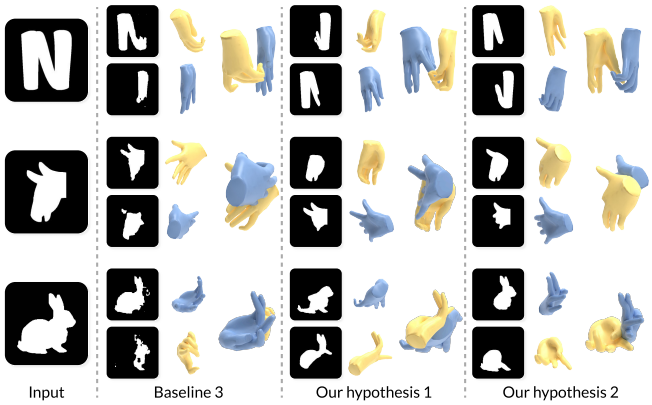


Fig. 17. Qualitative comparison between Hand-Shadow Poser and *Baseline 3*. Hand assignment results (left), coarse 3D hand poses (middle), and refined 3D hand poses (right) are shown for each case.

7.2 Evaluation

Gallery. We present our visual results for three classes of shapes: alphanumeric characters (C1) in Figure 11, real hand-shadow-art shapes (C2) in Figure 12, and shapes of diverse everyday objects (C3) in Figure 13, which exhibit varying levels of complexity. Details about the three classes of shapes are presented in Section 7.1. These results showcase the remarkable versatility of our method in reproducing many different kinds of object shapes, covering animals, plants, human portraits, logos, daily-used tools, numbers, letters, *etc.* More results are provided in the supplementary material.

Human demonstration. Next, we present some real shadow results. Figure 14 shows our physical setup and some example real shadows produced by human hands using a spotlight, demonstrating the feasibility of our method in practical scenarios.

3D fabrication. Further, we 3D-printed several 3D hand-poses results at a scale of 1:4 relative to the size of normal human hands. In detail, we printed a horizontal tube invisible from the front view to join the two disconnected hands and another vertical/L-shaped tube from the bottom to the middle of the horizontal tube to support the two-hand sculpture. Figure 15 shows the results. Interestingly, these 3D-printed hands look like simple hands in the real world, but if we look at them from a specific angle or shine a light in this direction, we can observe the shapes hidden by the hand sculptures.

Qualitative comparison. In Figure 16, we visually compare our method with the baselines on three target shadows: F, Flower, and Elephant. Our method can create high-fidelity results that adhere to the input shadow details, highlighted by successfully preserving the key characteristics of shape, including all the petals in Flower and the bending nose of the Elephant. Though *Baseline 3* can provide a relatively better initialization than the other baselines (based on the relative hand positions in Flower), it still struggles to produce a well-aligned shadow, primarily due to the segmentation model's ignorance of learning hand shapes from a global view.

Generative vs. *segmentation.* To further evaluate our generative approach against the segmentation-based approach, we compare the hand assignment results and recovered 3D hand poses before/after refinement from our method with those in *Baseline 3*. From Figure 17, we can see that our method is capable of producing diverse hand shapes of higher quality, in terms of smoothness and completeness in the hand masks, whereas the segmentation model in *Baseline 3* fails for complex input shapes like the Stanford bunny (see the over-segmentation artifacts). Consequently, under the same alignment and optimization process, our method yields multiple 3D hand poses with superior alignment to the target shadows, demonstrating the benefits of our generative formulation.

Quantitative comparison. We compare the three baselines with our method on all three classes of shapes in the benchmark. Table 1 reports the full results on seven metrics, including the five metrics presented in Section 7.1 and two human-related metrics to be presented later in the user study. Our method achieves the best results for all metrics and classes, showing its superiority in preserving the

Table 2. Mean iteration number and runtime for optimization convergence.

Methods	# Iterations ↓			Time (in seconds) ↓		
	C1 / C2 / C3		Avg.	C1 / C2 / C3		Avg.
Baseline 1	5809 / 5328 / 4792		5314	348 / 319 / 287		318
Baseline 2	5532 / 4883 / 3691		4728	331 / 293 / 221		283
Baseline 3	1845 / 1933 / 1705		1841	110 / 116 / 102		110
Ours	**1216** / **548** / **1234**		**945**	**73** / **32** / **74**		**56**

Table 3. Ablation study on the shadow-feature-aware refinement module.

Methods	*LPIPS* ↓	*CLIP-Global* ↑	*CLIP-Semantic* ↑	*DINO-Global* ↑	*DINO-Semantic* ↓
w/o refinement	0.19	0.92	0.30	0.69	0.76
w/o saliency	0.15	0.93	0.36	0.74	0.59
Full (ours)	**0.14**	**0.94**	**0.40**	**0.78**	**0.56**

shape and features in the target shadows for both perceptual and semantic similarity.

Runtime comparison. Our Hand-Shadow Poser includes two stages of feed-forward models and a test-time optimization; therefore, we report their running efficiency separately. For fairness, all runtime measurements were taken on a single NVIDIA RTX 2080Ti GPU.

First, the models in the previous two stages have an average processing time of 3 minutes and 30 milliseconds per shape.

For Stage 3, as discussed in [Hospedales et al. 2021], initialization plays a crucial role in preventing being stuck at local minima during the optimization. To show the advantages of our initialization from feed-forward models, we report the number of iterations and runtime for the optimization to converge in each baseline and our method. Specifically, the terminating condition is empirically set as (i) when the result's quality exceeds a certain *LPIPS* score calculated as the mean *LPIPS* of all four methods reported in Table 1, or (ii) when the optimization reaches a maximum of 6,000 iterations.

In Table 2, we can observe a significantly reduced running time of our method, particularly around 1/6 compared to random initialization in *Baseline 1*. Also, our method requires only half the optimization time of *Baseline 3*, demonstrating the superiority of taking a generative approach to produce more effective initial configurations. Combined with the results in Section 7.2, it indicates that the coarse hand poses from our first two stages not only yield better-optimized hand shadows but also accelerate the convergence. The efficacy of initialization from our method lies in the generalizability and robustness offered by the generative hand assignment and the generalized hand-shadow alignment module.

7.3 User Studies

We conducted two user studies to assess human preferences of the results produced by our approach versus the baselines, and also how our approach assists humans in creating hand shadow art.

Participants. We invited 13 volunteers aged 22 to 30 with no professional experience in hand shadow art. The participants are divided into two groups. The first group has 10 participants (5 males and 5 females) who helped to assess the visual quality of the rendered shadows and to perform human demonstrations. The other group (2 males and 1 female) served as judges in the second study to score the quality of human demonstrations after a brief tutorial session.

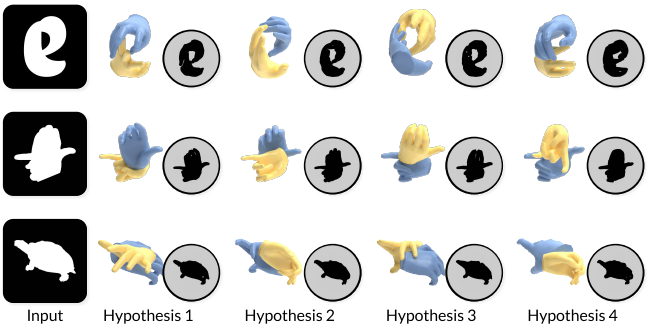
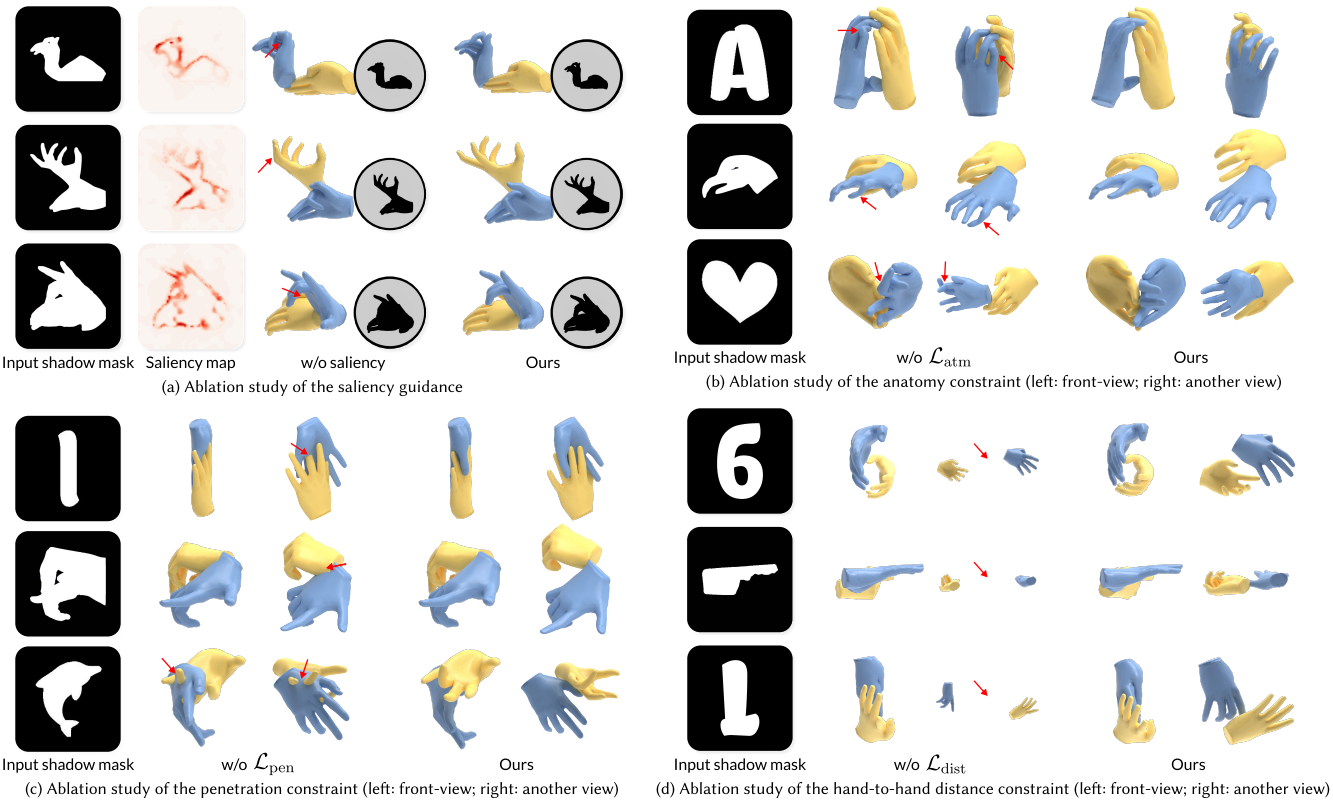


Fig. 18. Result diversity brought by our Hand-Shadow Poser. Note that the uniqueness is not limited to mirror symmetry, see results in the last row.

Metrics. For quality comparison with the baselines, the rendered shadows are evaluated in two aspects: (i) global shape similarity, which measures how similar the reproduced shadows are compared with the target shadows (*Human-Global*); and (ii) local details similarity, which measures the details preservation (*Human-Semantic*). The metrics are evaluated in a Likert scale from 1 (worst) to 5 (best). For the human demonstrations, we employ *Human-Semantic* to measure the quality of the reproduced real shadows. Also, we record the time taken by the participants in reproducing each shadow.

User study on quality comparison. Procedure-wise, we showed each participant 60 sets (20 shapes, for each shape class) of results from the three baselines and our method. The order of the results from different methods is randomly shuffled, with the associated target shape fixed on the left. For each result, we asked the participant to rate it on *Human-Global* and *Human-Semantic* by comparing it with the target shape. Table 1 (right) reports the average scores for each class. For all three classes, our method consistently achieves better scores in *Human-Global* and *Human-Semantic* than the baselines, confirming the satisfying perceptual quality of our results.

User study on human demonstration. In this study, the participants had to employ the physical setup described in Section 7.2 to recreate six shadows using their hands: Wolf, Sheep, Panther, Kangaroo, Dinosaur, and Stalin. In detail, we randomly and evenly split the ten participants in the first group into two sub-groups: the first sub-group aimed to reproduce the target shadows simply by taking the target shadow images as references, whereas the second sub-group performed the same task but additionally took our method's generated 3D hand models as references. Then, the three judges in the second group scored the quality of the reproduced shadows, following the *Human-Semantic* metric. Besides, we recorded the time taken to reach the best hand shadow by watching the video recordings of the reproducing procedures. Specifically, we recorded the timestamp of the best frame during the entire shadow-reproducing procedure for each shape within 3 minutes. Overall, our Hand-Shadow Poser helps reduce the average time taken by the participants, from 121.4



(a) Ablation study of the saliency guidance

(b) Ablation study of the anatomy constraint (left: front-view; right: another view)

(c) Ablation study of the penetration constraint (left: front-view; right: another view)

(d) Ablation study of the hand-to-hand distance constraint (left: front-view; right: another view)

Fig. 19. Ablation study of the key components in our shadow-feature-aware refinement module.

to 65.8 seconds, and improves the quality of the reproduced shadows, from 2.4 to 4.0 (on average).

7.4 Model Analysis

Ablation studies. Beyond comparisons with baselines, we additionally conduct ablation on the shadow-feature-aware refinement module in our pipeline, including removing (i) the whole refinement module, (ii) the saliency guidance in similarity constraint, (iii) the anatomy constraint, (iv) the penetration constraint, and (v) the hand-to-hand distance constraint, from our full design.

For cases (i) and (ii), we first provide a quantitative analysis in Table 3. The *DINO-Semantic* score drops significantly after removing the refinement module, showing the importance of shadow-specific optimization in achieving fine-grained shadow alignment. Though the effect of the saliency guidance is not obvious in *LPIPS*, for which we speculate is insensitive to local characteristics, the *DINO-Global* and *CLIP-Semantic* both show a moderate performance degradation without the saliency map. Besides, given the same initial pose for refinement, the visual ablation in Figure 19 (a) clearly shows the impact of the saliency guidance in preserving prominent and intricate details, such as the eyes of Camel and Donkey.

As the remaining three constraints in cases (iii-v) are directly imposed on the 3D hand poses to aim for physical plausibility, we show their visual ablation results, *i.e.*, 3D hand poses in Figure 19 (b-d). Comparing the areas highlighted with the red arrows, we can observe severe physical artifacts of the hands, including poor anatomy (Figure 19 (b)), penetration (Figure 19 (c)), and excessive distance ((Figure 19 (d))), after removing each constraint, manifesting the effectiveness of each of the associated constraint.

Diversity analysis. Further, we showcase multiple diverse results obtained by our method in Figure 18. For each case, we show four unique solutions, with the projected shadows closely resembling the input, while also remaining physically feasible. This manifests the diversity with reasonable hand shapes introduced by our carefully-designed generative hand assignment module.

Robustness analysis. Lastly, we study the robustness of our generalized hand-shadow alignment module. Figure 17 shows the coarse 3D hand poses of each hand without refinement (middle column in each case). For *all assignment results*, the coarse hand poses exhibit contours, positions, and orientations that are approximately consistent with the input hand shapes, revealing our method's robustness to handle input masks of varying levels of uncertainty, particularly evident in the Stanford bunny case of *Baseline 3*.

Failure case analysis. Hand-Shadow Poser may not be able to produce reasonable results for arbitrary inputs, as not all shapes can be effectively reproduced by hands, particularly those with intricate

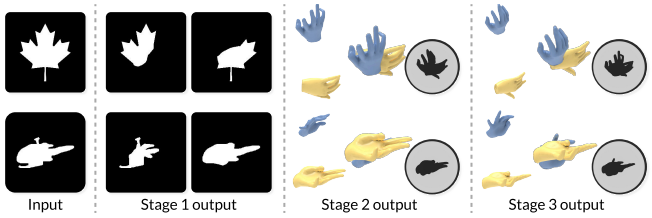


Fig. 20. Failure cases (Maple and Chopper). Our method may not be able to work on arbitrary inputs with intricate details or thin structures.

details or thin structures; see Figure 20. In such instances, our generative hand assignment may struggle to produce plausible hand shapes in Stage 1, thus leading to suboptimal hand reconstructions in Stage 2. Further, due to poor initialization, the inherent limitations of hand anatomy make it challenging for Stage 3 to refine poses to adequately fit the inputs.

8 CONCLUSION, LIMITATIONS, AND FUTURE WORKS

We presented the first comprehensive framework, namely Hand-Shadow Poser, to inversely create hand shadow arts from 2D shape inputs. We showcase the application of our approach to a wide variety of shapes, ranging from numbers and letters to classical hand shadows, and more challenging shapes of everyday objects. We contribute three notable advances: (i) first attempt to learn and reproduce hand shadow arts in a data-driven manner; (ii) a three-stage pipeline to decouple the anatomical constraints imposed by hand and semantic constraints imposed by shadow shape, with three novel components: generative hand assignment, generalized hand-shadow alignment, and shadow-feature-aware refinement. This decoupling design also frees us from building extensive domain-specific training data; and (iii) an evaluation benchmark with a rich variety of shadow art samples of varying complexity, along with a family of metrics for quantitative assessment. Also, we demonstrate the superior performance of our approach through extensive quantitative and qualitative comparisons with several alternative baselines and through user studies to evaluate human perception. In the end, we performed a series of analyses, including ablation on key components, result diversity, and model robustness, to study the effectiveness of our proposed designs.

Overall, the evaluation results highlight the generalizability and robustness of Hand-Shadow Poser in creating hand shadow arts with prominent features preserved for various types of input shapes, which can be further reproduced by human hands and 3D printing.

Limitations. The inverse hand-shadow-art problem is intriguing yet challenging. Our work still has some limitations. First, given an overly complicated shadow, such as shapes with small and thin structures, our approach may not be able to reproduce the shape due to the limited feasibility of human hands (see Section 7.4). Second, our approach cannot be directly applied to human hands of an arbitrary individual, due to variations in finger length and hand size, requiring customization by first specifying the hand-shape parameters of the individual. Third, we assume a fixed light source, which might not be feasible for some target shapes that are formed by distorted shadows. Fourth, a critical challenge is to ensure the feasibility of humans, since not all two-hand poses are achievable due to the anatomical limits of the human body, which cannot be resolved merely through physical constraints on hands. Last, our approach does not consider the forearm, which cannot be neglected in practice, as it may obscure the contour of the hand shadows. To incorporate the forearm into our pipeline, a potential solution is to extend the MANO hand model with forearm parameters (*e.g.*, via SMPL-X [Pavlakos et al. 2019]) for pose optimization, with a penalty term in Stage 3 to deviate the forearm shadows from obstructing critical hand features. Additionally, the feasibility issue can be partially addressed by restricting the left/right hand swapping based on the arm position constraints.

Future Works. Currently, our approach focuses on two hands to create hand shadow art from a single target shadow. First, we are interested in extending our approach to designing animated hand shadow arts for storytelling. Second, it would be interesting to incorporate hand-held object(s) in our approach, so that we may produce more intricate and visually appealing shadow results. Lastly, given that some shadow plays involve more than two hands from multiple artists, it would be intriguing to adapt Hand-Shadow Poser to coordinate the hands of multiple humans, enabling more elaborate and captivating hand shadow art creation.

ACKNOWLEDGMENTS

This work was partially supported by the Research Grants Council of the Hong Kong Special Administrative Region, China, under Project T45-401/22-N and No. CUHK 14201321. Niloy J. Mitra was partially supported by gifts from Adobe and the UCL AI Centre. Hao Xu thanks for the care and support from Yutong Zhang and his family.

REFERENCES

Ankur Agarwal and Bill Triggs. 2004. 3D human pose from silhouettes by relevance vector regression. In *CVPR*, Vol. 2. IEEE, II–II.

Marc Alexa and Wojciech Matusik. 2010. Reliefs as images. *ACM Transactions on Graphics (SIGGRAPH 2010)* 29, 4 (July 2010). https://doi.org/10.1145/1778765.1778797

Marc Alexa and Wojciech Matusik. 2012. Irregular pit placement for dithering images by self-occlusion. *Computers & Graphics* 36, 6 (2012), 635–641.

Y. ill Almoznino, Albert; Pinas. 1970. *The art of hand shadows.* Stravon Educational Press, New York.

Luca Ballan, Aparna Taneja, Jürgen Gall, Luc Van Gool, and Marc Pollefeys. 2012. Motion capture of hands in action using discriminative salient points. In *ECCV*. Springer, 640–653.

Ilya Baran, Philipp Keller, Derek Bradley, Stelian Coros, Wojciech Jarosz, Derek Nowrouzezahrai, and Markus Gross. 2012. Manufacturing layered attenuators for multiple prescribed shadow images. In *Computer Graphics Forum*, Vol. 31. Wiley Online Library, 603–610.

Amit Bermano, Ilya Baran, Marc Alexa, and Wojciech Matusk. 2012. SHADWOPIX: Multiple images from self shadowing. In *Computer Graphics Forum*, Vol. 31. Wiley Online Library, 593–602.

Alexandre Binninger and Olga Sorkine-Hornung. 2024. SD-πXL: Generating low-resolution quantized imagery via score distillation. In *ACM SIGGRAPH Asia 2024 Conference Papers*. 1–12.

Blender 2019. Blender. http://www.blender.org.

Ryan Burgert, Xiang Li, Abe Leite, Kanchana Ranasinghe, and Michael Ryoo. 2024. Diffusion illusions: Hiding images in plain sight. In *ACM SIGGRAPH 2024 Conference Papers*. 1–11.

Kartik Chandra, Tzu-Mao Li, Joshua Tenenbaum, and Jonathan Ragan-Kelley. 2022. Designing perceptual puzzles by differentiating probabilistic programs. In *ACM SIGGRAPH 2022 Conference Proceedings*. 1–9.

Li-Jen Chang, Yu-Cheng Liao, Chia-Hui Lin, Shih-Fang Yang-Mao, and Hwann-Tzong Chen. 2023. Mask2Hand: Learning to predict the 3D hand pose and shape from

shadow. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 591–598.
Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. 2021. DexYCB: A benchmark for capturing hand grasping of objects. In *CVPR*. 9044–9053.
Xiaozhong Chen, Sheldon Andrews, Derek Nowrouzezahrai, and Paul G. Kry. 2017. Ballistic shadow art. In *Proceedings of the 43rd Graphics Interface Conference (GI '17)*. Canadian Human-Computer Communications Society, Waterloo, CAN, 190–198.
Xingyu Chen, Yufeng Liu, Yajiao Dong, Xiong Zhang, Chongyang Ma, Yanmin Xiong, Yuan Zhang, and Xiaoyan Guo. 2022. MobRecon: Mobile-friendly hand mesh reconstruction from monocular image. In *CVPR*. 20544–20554.
Ming-Te Chi, Tong-Yee Lee, Yingge Qu, and Tien-Tsin Wong. 2008. Self-animating images: Illusory motion using repeated asymmetric patterns. *ACM Transactions on Graphics (TOG)* 27, 3 (2008), 1–8.
Chun-Chia Chiu, Yi-Hsiang Lo, Ruen-Rone Lee, and Hung-Kuo Chu. 2015. Tone- and feature-aware circular scribble art. In *Computer Graphics Forum*, Vol. 34. Wiley Online Library, 225–234.
Hung-Kuo Chu, Wei-Hsin Hsu, Niloy J. Mitra, Daniel Cohen-Or, Tien-Tsin Wong, and Tong-Yee Lee. 2010. Camouflage images. *ACM Transactions on Graphics* 29, 4 (2010), 51–1.
J. Coren, S.; Girgus. 1978. *Seeing is deceiving: The psychology of visual illusions*. Routledge, London.
Endri Dibra, Himanshu Jain, Cengiz Oztireli, Remo Ziegler, and Markus Gross. 2017. Human shape from silhouettes using generative hks descriptors and cross-modal neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.* 4826–4836.
Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. 2022. AlphaPose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 6 (2022), 7157–7173.
Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*. PMLR, 1126–1135.
Ran Gal, Olga Sorkine, Tiberiu Popa, Alla Sheffer, and Daniel Cohen-Or. 2007. 3D collage: Expressive non-realistic modeling. In *NPAR: Proceedings of the 5th International Symposium on Non-Photorealistic Animation and Rendering*. ACM Press, 7–14.
Aalok Gangopadhyay, Paras Gupta, Tarun Sharma, Prajwal Singh, and Shanmuganathan Raman. 2024. Search me knot, render me knot: Embedding search and differentiable rendering of knots in 3D. *Computer Graphics Forum* 43, 5 (August 2024), i–x.
Aalok Gangopadhyay, Prajwal Singh, Ashish Tiwari, and Shanmuganathan Raman. 2023. Hand shadow art: A differentiable rendering perspective. In *Pacific Graphics Short Papers and Posters*. The Eurographics Association. https://doi.org/10.2312/pg.20231279
Gekidan Kakashiza 1952. Shadow Play Theatre KAKASHIZA. https://kakashiza-en.com/.
Daniel Geng, Inbum Park, and Andrew Owens. 2024. Visual anagrams: Generating multi-view optical illusions with diffusion models. In *CVPR*. 24154–24163.
Daniel Geng, Inbum Park, and Andrew Owens. 2025. Factorized diffusion: Perceptual illusions by noise decomposition. In *ECCV*. Springer, 366–384.
Francisco Gomez-Donoso, Sergio Orts-Escolano, and Miguel Cazorla. 2019. Large-scale multiview 3D hand pose dataset. *Image and Vision Computing* 81 (2019), 25–33.
Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. 2020. Honnotate: A method for 3D annotation of hand and object poses. In *CVPR*. 3196–3206.
Ayaan Haque, Matthew Tancik, Alexei A. Efros, Aleksander Holynski, and Angjoo Kanazawa. 2023. Instruct-NeRF2NeRF: Editing 3D scenes with instructions. In *ICCV*. 19740–19750.
Ryuji Hirayama, Hirotaka Nakayama, Atushi Shiraki, Takashi Kakue, Tomoyoshi Shimobaba, and Tomoyoshi Ito. 2019. Projection of multiple directional images on a volume structure with refractive surfaces. *Optics Express* 27, 20 (2019), 27637–27648.
Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *NeurIPS*, Vol. 33. 6840–6851.
Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).
Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. 2021. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence* 44, 9 (2021), 5149–5169.
Seyed Vahab Hosseini, Usman Alim, Ali Mahdavi Amiri, Lora Oehlberg, and Joshua Taron. 2020. Portal: Design and fabrication of incidence-driven screens. *International society of the arts, mathematics, and architecture, summer* (2020), 31–46.
Kai-Wen Hsiao, Jia-Bin Huang, and Hung-Kuo Chu. 2018. Multi-view wire art. *ACM Transactions on Graphics* 37, 6 (2018), 242.
Lin Huang, Chung-Ching Lin, Kevin Lin, Lin Liang, Lijuan Wang, Junsong Yuan, and Zicheng Liu. 2023. Neural voting field for camera-space 3D hand pose estimation. In *CVPR*. 8969–8978.
Umar Iqbal, Pavlo Molchanov, Thomas Breuel, Juergen Gall, and Jan Kautz. 2018. Hand pose estimation via latent 2.5D heatmap regression. In *ECCV*. 118–134.
Frank Jacobs. 1996. *Fun with hand shadows*. Dover Publications, Mineola, N.Y.
Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. 2021. Hand-object contact consistency reasoning for human grasps generation. In *ICCV*. 11107–11116.
Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. 2020. Whole-body human pose estimation in the wild. In *ECCV*. Springer, 196–214.
Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. 2022. Elucidating the design space of diffusion-based generative models. *NeurIPS* 35 (2022), 26565–26577.
JJA Keiren, Freek van Walderveen, and Alexander Wolff. 2009. Constructability of triplets. In *Abstracts 25th European Workshop on Computational Geometry (EuroCG'09, Brussels, Belgium, March 16-18, 2009)*. 251–254.
Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization.. In *ICLR (Poster)*. http://dblp.uni-trier.de/db/conf/iclr/iclr2015.html#KingmaB14
Johannes Kopf and Dani Lischinski. 2011. Depixelizing pixel art. In *ACM SIGGRAPH 2011 papers*. 1–8.
Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet classification with deep convolutional neural networks. *NeurIPS* 25 (2012).
Hae Beom Lee, Hayeon Lee, Donghyun Na, Saehoon Kim, Minseop Park, Eunho Yang, and Sung Ju Hwang. 2020. Learning to balance: Bayesian meta-learning for imbalanced and out-of-distribution tasks. In *ICLR*.
Kuo-Wei Lee, Shih-Hung Liu, Hwann-Tzong Chen, and Koichi Ito. 2019. Silhouette-Net: 3D hand pose estimation from silhouettes. *arXiv preprint arXiv:1912.12436* (2019).
Lijun Li, Linrui Tian, Xindi Zhang, Qi Wang, Bang Zhang, Liefeng Bo, Mengyuan Liu, and Chen Chen. 2023. RenderIH: A large-scale synthetic dataset for 3D interacting hand pose estimation. In *ICCV*. 20395–20405.
Juncong Lin, Takeo Igarashi, Jun Mitani, Minghong Liao, and Ying He. 2012. A sketching interface for sitting pose design in the virtual environment. *IEEE transactions on visualization and computer graphics* 18, 11 (2012), 1979–1991.
Kunhao Liu, Fangneng Zhan, Muyu Xu, Christian Theobalt, Ling Shao, and Shijian Lu. 2024. StyleGaussian: Instant 3D style transfer with gaussian splatting. In *SIGGRAPH Asia 2024 Technical Communications*. 1–4.
Lingjie Liu, Duygu Ceylan, Cheng Lin, Wenping Wang, and Niloy J. Mitra. 2017. Image-based reconstruction of wire art. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–11.
Xinyu Liu, Houwen Peng, Ningxin Zheng, Yuqing Yang, Han Hu, and Yixuan Yuan. 2023. EfficientViT: Memory efficient vision transformer with cascaded group attention. In *CVPR*. 14420–14430.
Oliver Mattausch, Takeo Igarashi, and Michael Wimmer. 2013. Freeform shadow boundary editing. In *Computer Graphics Forum*, Vol. 32. Wiley Online Library, 175–184.
Wojciech Matusik, Boris Ajdin, Jinwei Gu, Jason Lawrence, Hendrik P. A. Lensch, Fabio Pellacini, and Szymon Rusinkiewicz. 2009. Printing spatially-varying reflectance. *ACM Transactions on Graphics (TOG)* 28, 5 (2009), 1–9.
Sehee Min, Jaedong Lee, Jungdam Won, and Jehee Lee. 2017. Soft shadow art. In *Proceedings of the symposium on Computational Aesthetics*. 1–9.
Niloy J. Mitra and Mark Pauly. 2009. Shadow art. *ACM Transactions on Graphics (TOG)* 28, 5 (2009), 156–1.
Gyeongsik Moon and Kyoung Mu Lee. 2020. I2L-MeshNet: Image-to-lixel prediction network for accurate 3D human pose and mesh estimation from a single RGB image. In *ECCV*. 752–768.
Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. 2020. InterHand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image. In *ECCV*. Springer, 548–564.
Louis Nikola. 1913. *Hand shadows: The complete art of shadowgraphy*. C. Arthur Pearson, LTD., Henrietta Street, W.C.
Aude Oliva, Antonio Torralba, and Philippe. G. Schyns. 2006. Hybrid images. *ACM Transactions on Graphics (TOG)* 25, 3 (2006), 527–532.
Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).
Marios Papas, Thomas Houit, Derek Nowrouzezahrai, Markus H. Gross, and Wojciech Jarosz. 2012. The magic lens: Refractive steganography. *ACM Transactions on Graphics* 31, 6 (2012), 186–1.
Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. PyTorch: An imperative style, high-performance deep learning library. *NeurIPS* 32 (2019).
Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*.
Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. 2024. Reconstructing hands in 3D with transformers. In *CVPR*. 9826–9836.

Fabio Pellacini, Parag Tole, and Donald P. Greenberg. 2002. A user interface for interactive cinematic shadow design. *ACM Transactions on Graphics (TOG)* 21, 3 (2002), 563–566.

Maxine Perroni-Scharf and Szymon Rusinkiewicz. 2023. Constructing printable surfaces with view-dependent appearance. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–10.

Zhiyu Qu, Lan Yang, Honggang Zhang, Tao Xiang, Kaiyue Pang, and Yi-Zhe Song. 2024. Wired perspectives: Multi-view wire art embraces generative AI. In *CVPR*. 6149–6158.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. 2019. Meta-learning with implicit gradients. *NeurIPS* 32 (2019).

Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2012. Reconstructing 3D human pose from 2D image landmarks. In *ECCV*. Springer, 573–586.

Javier Romero, Dimitrios Tzionas, and Michael J Black. 2022. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610* (2022).

Kaustubh Sadekar, Ashish Tiwari, and Shanmuganathan Raman. 2022. Shadow art revisited: A differentiable rendering based approach. In *WACV*. 29–37.

Kaisei Sakurai, Yoshinori Dobashi, Kei Iwasaki, and Tomoyuki Nishita. 2018. Fabricating reflectors for displaying multiple images. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–10.

Christian Schüller, Daniele Panozzo, and Olga Sorkine-Hornung. 2014. Appearance-mimicking surfaces. *ACM Transactions on Graphics (TOG)* 33, 6 (2014), 1–10.

Guy Sela and Gershon Elber. 2007. Generation of view dependent models using free form deformation. *The Visual Computer* 23, 3 (2007), 219–229.

Pengfei Shen, Rui-Zeng Li, Beibei Wang, and Ligang Liu. 2023. Scratch-based reflection art via differentiable rendering. *ACM Transactions on Graphics* 42, 4 (2023), 65–1.

Xiao Shen. 2024. Puppet and shadow dramas. http://www.cdsfyy.com/portal/article/index.html?id=45&cid=14.

Thomas Sikora. 2001. The MPEG-7 visual standard for content description-an overview. *IEEE Transactions on circuits and systems for video technology* 11, 6 (2001), 696–702.

Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. 2017. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*. 1145–1153.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

Kokichi Sugihara. 2014. Design of solids for antigravity motion illusion. *Computational Geometry* 47, 6 (2014), 675–682.

Kenji Tojo, Ariel Shamir, Bernd Bickel, and Nobuyuki Umetani. 2024. Fabricable 3D wire art. In *ACM SIGGRAPH 2024 Conference Papers*. 1–11.

Jing Tong, Ligang Liu, Jin Zhou, and Zhigeng Pan. 2013. Mona Lisa alive: Create self-moving objects using hollow-face illusion. *The Visual Computer* 29 (2013), 535–544.

Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. 2016. Capturing hands in action using discriminative salient points and physics simulation. *International Journal of Computer Vision (IJCV)* 118, 2 (June 2016), 172–193. https://doi.org/10.1007/s11263-016-0895-4

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*, Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Bingyuan Wang, Qifeng Chen, and Zeyu Wang. 2024. Diffusion-based visual art creation: A survey and new perspectives. *arXiv preprint arXiv:2408.12128* (2024).

Caoliwen Wang and Bailin Deng. 2024. Neural shadow art. (2024). arXiv:2411.19161 [cs.CV] https://arxiv.org/abs/2411.19161

Tim Weyrich, Pieter Peers, Wojciech Matusik, and Szymon Rusinkiewicz. 2009. Fabricating microgeometry for custom surface reflectance. *ACM Transactions on Graphics (TOG)* 28, 3 (2009), 1–6.

Jungdam Won and Jehee Lee. 2016. Shadow theatre: Discovering human motion from a sequence of silhouettes. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 1–12.

Kang Wu, Renjie Chen, Xiao-Ming Fu, and Ligang Liu. 2022. Computational mirror cup and saucer art. *ACM Transactions on Graphics (TOG)* 41, 5 (2022), 1–15.

Tai-Pang Wu, Chi-Wing Fu, Sai-Kit Yeung, Jiaya Jia, and Chi-Keung Tang. 2010. Modeling and rendering of impossible figures. *ACM Transactions on Graphics (TOG)* 29, 2 (2010), 1–15.

Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. 2019. Monocular total capture: Posing face, body, and hands in the wild. In *CVPR*. 10965–10974.

Hao Xu, Tianyu Wang, Xiao Tang, and Chi-Wing Fu. 2023. H2ONet: Hand-occlusion-and-orientation-aware network for real-time 3D hand mesh reconstruction. In *CVPR*. 17048–17058.

Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. 2021b. CPF: Learning a contact potential field to model the hand-object interaction. In *ICCV*.

Zhijin Yang, Pengfei Xu, Hongbo Fu, and Hui Huang. 2021a. WireRoom: Model-guided explorative design of abstract wire art. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–13.

Yonghao Yue, Kei Iwasaki, Bing-Yu Chen, Yoshinori Dobashi, and Tomoyuki Nishita. 2012. Pixel art with refracted light by rearrangeable sticks. In *Computer Graphics Forum*, Vol. 31. Wiley Online Library, 575–582.

Jiani Zeng, Honghao Deng, Yunyi Zhu, Michael Wessely, Axel Kilian, and Stefanie Mueller. 2021. Lenticular objects: 3D printed objects with lenticular lens surfaces that can change their appearance depending on the viewpoint. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. 1184–1196.

Jiawei Zhang, Jianbo Jiao, Mingliang Chen, Liangqiong Qu, Xiaobin Xu, and Qingxiong Yang. 2017b. A hand pose tracking benchmark from stereo matching. In *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 982–986.

Kai Zhang, Nick Kolkin, Sai Bi, Fujun Luan, Zexiang Xu, Eli Shechtman, and Noah Snavely. 2022. ARF: Artistic radiance fields. In *ECCV*. Springer, 717–733.

Qing Zhang, Gelin Yin, Yongwei Nie, and Wei-Shi Zheng. 2020. Deep camouflage images. In *AAAI*, Vol. 34. 12845–12852.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*.

Xiong Zhang, Hongsheng Huang, Jianchao Tan, Hongmin Xu, Cheng Yang, Guozhu Peng, Lei Wang, and Ji Liu. 2021. Hand image understanding via deep multi-task learning. In *ICCV*. 11281–11292.

Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. 2019. End-to-end hand mesh recovery from a monocular RGB image. In *ICCV*. 2354–2364.

Yanxiang Zhang, Dong Dong, and Yanlong Guo. 2017a. 3D shadow art sculpture based on real items. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition*. 243–246.

Haisen Zhao, Lin Lu, Yuan Wei, Dani Lischinski, Andrei Sharf, Daniel Cohen-Or, and Baoquan Chen. 2016. Printed perforated lampshades for continuous projective images. *ACM Transactions on Graphics (TOG)* 35, 5 (2016), 1–11.

Pancheng Zhao, Peng Xu, Pengda Qin, Deng-Ping Fan, Zhicheng Zhang, Guoli Jia, Bowen Zhou, and Jufeng Yang. 2024. LAKE-RED: Camouflaged images generation by latent background knowledge retrieval-augmented diffusion. In *CVPR*. 4092–4101.

Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. 2020. Monocular real-time hand shape and motion capture using multi-modal data. In *CVPR*. 5346–5355.

Zhishan Zhou, Shihao Zhou, Zhi Lv, Minqiang Zou, Yao Tang, and Jiajun Liang. 2024. A simple baseline for efficient hand mesh reconstruction. In *CVPR*. 1367–1376.

Amy Zhu, Yuxuan Mei, Benjamin Jones, Zachary Tatlock, and Adriana Schulz. 2024. Computational illusion knitting. *ACM Transactions on Graphics (TOG)* 43, 4 (2024), 1–13.

Christian Zimmermann and Thomas Brox. 2017. Learning to estimate 3D hand pose from single RGB images. In *ICCV*. 4903–4911.

Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. 2019. FreiHAND: A dataset for markerless capture of hand pose and shape from single RGB images. In *ICCV*. 813–822.

Binghui Zuo, Zimeng Zhao, Wenqian Sun, Wei Xie, Zhou Xue, and Yangang Wang. 2023. Reconstructing interacting hands with interaction prior from monocular images. In *ICCV*. 9054–9064.