Creating a Forensic Database of Shoeprints from Online Shoe-Tread Photos

Samia Shafique[1] Bailey Kong[2] Shu Kong[3]* Charless Fowlkes[1]*
[1]University of California, Irvine [2]Ronin Institute [3]Texas A&M University
sshafiqu@uci.edu bailey.kong@ronininstitute.org fowlkes@ics.uci.edu shu@tamu.edu
https://github.com/Samia067/ShoeRinsics

Abstract

Shoe-tread impressions are one of the most common types of evidence left at crime scenes. However, the utility of such evidence is limited by the lack of databases of footwear prints that cover the large and growing number of distinct shoe models. Moreover, the database is preferred to contain the 3D shape, or depth, of shoe-tread photos so as to allow for extracting shoeprints to match a query (crime-scene) print. We propose to address this gap by leveraging shoe-tread photos collected by online retailers. The core challenge is to predict depth maps for these photos. As they do not have ground-truth 3D shapes allowing for training depth predictors, we exploit synthetic data that does. We develop a method, termed **ShoeRinsics**, *that learns to predict depth from fully supervised synthetic data and unsupervised retail image data. In particular, we find domain adaptation and intrinsic image decomposition techniques effectively mitigate the synthetic-real domain gap and yield significantly better depth predictions. To validate our method, we introduce 2 validation sets consisting of shoe-tread image and print pairs and define a benchmarking protocol to quantify the quality of predicted depth. On this benchmark, ShoeRinsics outperforms existing methods of depth prediction and synthetic-to-real domain adaptation.*

1. Introduction

Studying the evidence left at a crime scene aids investigators in identifying criminals. Shoeprints have a greater chance of being present at crime scenes [9], although they may have fewer uniquely identifying characteristics than other biometric samples (such as blood or hair). Thus, studying shoeprints can provide valuable clues to help investigators narrow down suspects of a crime.

Forensic analysis of shoeprints can provide clues on the *class characteristics* and the *acquired characteristics* of the suspect's shoe. The former involves the type of shoe (e.g., the brand, model, and size); the latter consists of the individual traits of a particular shoe that appear over time as it is worn (e.g., holes, cuts, and scratches). We are interested in aiding the study of class characteristics of shoeprints.

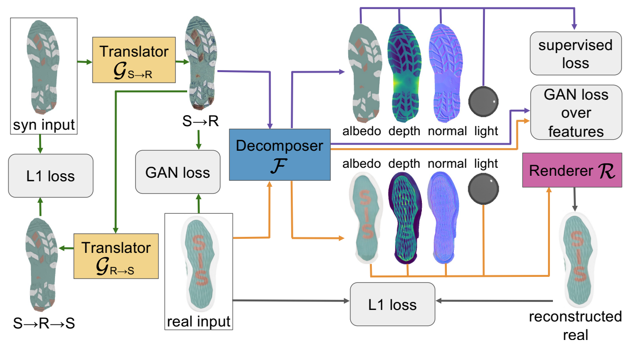


Figure 1: Predicting depth for shoe-tread images (collected by online retailers) is the core challenge in constructing a shoeprint database for forensic use. We develop a method termed *ShoeRinsics* to learn depth predictors. The flowchart depicts how we train *ShoeRinsics* using annotated synthetic and un-annotated real images (Sec. 4). We use domain adaptation (via image translators $\mathcal{G}_{S\to R}$ and $\mathcal{G}_{R\to S}$) and intrinsic image decomposition (via decomposer $\mathcal{F}$ and renderer $\mathcal{R}$) techniques to mitigate synthetic-real domain gaps (Sec. 5). Our method achieves significantly better depth prediction on real shoe-tread images than the prior art (Sec. 6).

Status quo. Traditionally, investigating class characteristics of shoeprints involve matching the prints against a *manually curated* database of impressions of various shoe models [11]. The research community has shown significant interest in automating this matching process [10, 12, 21, 22, 23, 3, 29, 33, 31, 57, 65]. However, in practice, the success of such work depends on the quality of the database to which the shoeprint evidence is compared. Yet, maintaining and regularly updating such a database to include all shoe models is tedious, costly, and requires significant human effort. Shoeprint matching methods are decidedly less useful if the database does not include the type of shoe the criminal wore! Partly because of this, shoeprint evidence is vastly underutilized in the USA [52].

Motivation. To address the need for such a comprehensive database, we propose to leverage imagery of shoe-treads collected by online retailers. High-resolution

*Authors share senior authorship.

tread photos of various shoe products are readily available, and shopping websites are updated frequently (>1000 new products appear each month based on our analysis on some websites). Fig. 2 (b) shows examples of such shoe-tread images. Developing a method to predict the 3D shape from a shoe-tread image would directly address the need for a comprehensive, up-to-date database of tread patterns. *We formulate this problem as depth prediction for shoe-treads; thresholding the depth map of a given shoe can generate/simulate shoeprints sufficient for matching query prints.*

Technical Insights. To learn depth predictors from single shoe-tread images, we would ideally utilize supervised training examples of aligned shoe-tread images and their corresponding depth maps. However, since such ground-truth data is simply unavailable, we develop an alternative strategy. We create a synthetic dataset of rendered shoe-tread images and corresponding ground-truth depth, albedo, normal, and lighting. This data can train a predictor in a fully supervised fashion. However, the resulting model performs sub-optimally on real-world images due to the domain gap between synthetic and real imagery. To address this, we introduce three additional techniques to close the synthetic-real domain gap by incorporating methods of domain adaptation [70] and intrinsic image decomposition [28] (see Fig. 1). First, we train a translator that translates synthetic shoe-treads to realistic images, which better match the distribution of the real shoe-treads. Second, we use an adversarial loss to enforce that features of real and translated synthetic images are indistinguishable. Third, we use a re-rendering loss that adopts a synthetically trained renderer to reconstruct the real shoe-tread images using their predicted depth and other intrinsic components. We find these three techniques in combination help close the domain gap and yield significantly better depth prediction.

Contributions. We make three major contributions.

- Motivated to create a database of shoeprints for forensic use, we introduce the task of depth prediction for real shoe-tread photos collected by online retailers.
- We develop a benchmarking protocol, with which we evaluate existing methods of depth prediction using domain adaptation for this task.
- We develop a method called *ShoeRinsics* that incorporates intrinsic image decomposition and domain adaptation techniques, outperforming prior art for this task.

2. Related Work

Shoeprint Analysis. Automatic shoeprint matching has been studied widely in the past two decades [45]. Existing works focus on generating good features from shoeprints and using them to assign a class label (shoe type) from a database of lab footwear impressions. To study global features (i.e., considering the whole shoe), [33] introduces a probabilistic compositional active basis model, [31] explores multi-channel normalized cross-correlation to match multi-channel deep features, and [57] employs a manifold ranking method, and [65] uses VGG16 as a feature extractor. On the other hand, [41] studies a multi-part weighted CNN, [5] introduces a block sparse representation technique, and [6] applies multiple point-of-interest detectors and SIFT descriptors to study the local features of shoeprints (i.e., keypoints [34]). Our work differs from the previous work as it focuses on creating a database of prints rather than developing methods for shoeprint matching. *Creating such as database is a prerequisite for algorithmic explorations for shoe-matching.*

Monocular Depth Prediction has been studied extensively since early works [26, 48, 47]. Previous methods invent features representations [8, 44, 18], deep network architectures [7, 37, 46, 30, 36, 60], and training losses [16, 51, 62]. [35, 20, 39] explore self-supervised learning in a stereo setup while [43, 67] experiment with training on large datasets. Depth estimation has been further improved by considering the camera pose [66]. Our work differs from the above as it aims for depth prediction on real images by learning over un-annotated real images and synthetic images (and their ground-truth intrinsics: depth, albedo, normal and light).

Intrinsic Image Decomposition. Another line of work aims to explain image appearance in terms of some intrinsic components, including albedo, normals, and lighting. However, predicting intrinsic images is difficult, if not impossible. Our approach is related to [28], which learns for intrinsic image decomposition and uses a differentiable renderer to leverage un-annotated images with a reconstruction loss. [50, 42, 58] focus on face images and explore a similar reconstruction loop [50], non-diffuse lighting models [42], and multiple reflectance channels [58]. [59] works on rotationally symmetric objects with only object silhouettes as supervision. [49, 64, 69, 38, 63] study decomposition on entire scenes. [4] learns photo-realistic rendering of synthetic data and intrinsic decomposition of real images using unpaired data as input via an adversarial loss. In contrast, our work utilizes intrinsic decomposition techniques to help learn depth prediction by leveraging annotated synthetic and un-annotated real data via domain adaptation.

Domain Adaptation. Training solely on synthetic data can cause models to perform poorly on real data. Adversarial domain adaptation has proved promising for bridging such domain gaps. One way to approach this is to use domain-invariant features to map between the domains. [40] proposes to reduce the Maximum Mean Discrepancy to learn domain-invariant features. [56] builds on this idea and further improves domain adaptation performance in classification tasks. [55, 54, 19, 53] learn domain adaptation by aligning source and target features. Another direction of work uses image-to-image translation [70] to stylize source

Table 1: Overview of our datasets for training and testing, along with their shoe categories and counts. It is worth noting that real-val contains formal and used shoes, which are not present in training (i.e., the real-train set). We include these novel shoe types to analyze the generalizability of different methods. See details in Sec. 4 and visual examples in Fig. 2.

Dataset	Shoe Category			Total	Annotation
	New-Athletic	Formal	Used		
syn-train	88,408	0	0	88,408	depth, albedo, normal, light
real-train	3,543	0	0	3,543	none
real-val	22	6	8	36	print
real-FID-val	41	0	0	41	print

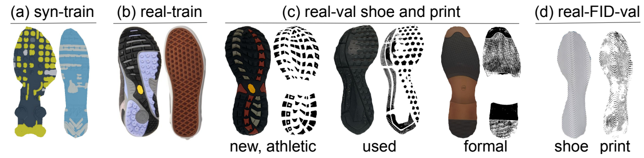


Figure 2: Shoe tread examples from (a) syn-train, (b) real-train, (c) real-val, and (d) real-FID-val. Clearly, a domain gap exists between (a) syn-train and (b) real-train, demonstrating the need to close the synthetic-real domain gap. Moreover, to study the generalizability, we evaluate on 2 datasets (c) and (d) and purposely hold out the formal and used shoe-treads which are not used for training but for validation (c).

images as target images. [25, 68] use the stylized source images to learn from target images using source labels while performing alignment both at the image and feature level. We use domain adaptation for depth estimation but take this approach further by reasoning about the intrinsic components of unlabeled real data.

3. Problem Setup and Evaluation Protocol

Our motivation is to create a database of shoeprints for forensic use. *The specific task is to predict depth maps for shoe-tread images collected by online retailers.* Below, we formulate the problem and introduce an evaluation protocol to benchmark methods.

3.1. Problem Setup

Online shoe-tread photos do not have ground-truth depth. Thus, we cannot directly train a depth predictor on them. Instead, we propose to create a dataset of synthetic shoe-tread images for which we have a complete set of annotations, including depth, albedo, normal, and lighting (details in Section 4.1). Therefore, **the problem is to predict depth for *real* shoe-treads by learning a depth predictor on synthetic shoe-treads (with annotations) and real shoe-treads (without annotations)**. This requires (1) learning a depth predictor by exploiting synthetic data that has annotations of depth and other intrinsic components, (2) addressing the synthetic-real domain gap.

3.2. Evaluation Protocol

Recall that the created database, containing predicted depth maps and shoe-tread images, and will serve for forensic use – an investigator will query a shoeprint collected at a crime scene by matching it with depth maps within this database. Therefore, we evaluate the quality of predicted depth maps w.r.t shoeprint matching.

To this end, we introduce two validation sets that contains paired "ground-truth" shoeprints and shoe-tread photos (details in Section 4.2). For a given shoe-tread, a trained model predicts its depth and the metric measures the degree of match between the ground-truth shoeprint and the predicted depth. We develop a metric based on Intersection-over-Union (IoU). Specifically, we generate a set of shoeprints using adaptive thresholding (with a range of hyperparameters) for the predicted depth, and compute the IoU between the ground-truth print to each of these generated shoeprints. The metric returns the highest IoU. We further average the IoUs over all the validation data as mean IoU (mIoU) to benchmark methods. Refer to the supplement for further details.

4. Data Preparation

During training, we have two data sources: a synthetic dataset (*syn-train*) that has annotations, and a dataset of unannotated real shoe-treads (*real-train*). To study models' generalizability, we test our model on two validation sets (*real-val* and *real-FID-val*). Each of these datasets contain shoe-tread photos with aligned ground-truth shoeprints, which enable quantitative evaluation. Note that to analyze the models' robustness to novel shoe types, we constrain our training sets to contain only brand-new athletic shoes while letting real-val also include formal and used (worn) shoes. Fig. 2 displays example shoe-treads and Table 1 summarizes the four datasets. Below, we elaborate on the creation of the synthetic training set (syn-train), the real training set (real-train), and validation sets (real-val and real-FID-val).

4.1. Synthetic Data for Training

Our synthetic dataset (syn-train) containing synthetic shoe-tread images and their intrinsic annotations (depth, albedo, normal, and lighting). We synthesize a shoe-tread image with a given depth map, an albedo map, and a lighting environment (outlined in Fig. 3). We pass these to a physically-based rendering engine [27] to generate the synthetic image. The final syn-train set contains 88,408 shoe-treads with paired ground-truth intrinsic images.

Depth Map. We use an existing dataset [61] to generate plausible synthetic depth maps to create syn-train data. For each of 387 shoeprints, we synthesize 10-15 different depth maps. Because the shoeprints have noise that affects synthetic data generation, we first apply a Gaussian blur to filter

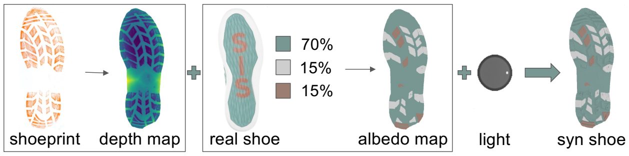


Figure 3: Generation of synthetic data. We scale off-the-shelf shoeprints to generate "pseudo" depth maps. We sample a color distribution from a real-shoe example to create an albedo map. The depth map and albedo map are combined with a lighting environment to render a synthetic image. The lighting environment is demonstrated by visualizing a shiny sphere in place of the shoe. In this example, directional light comes from a point on the right.

the noise. We then scale the blurred print image to create a "pseudo" depth map. To generate more diverse depth maps we add random high-frequency textures. Lastly, we make tread shapes more realistic by adding a priori features, such as slanted bevels on the tread elements and global curvature of the shoe-tread (details in supplement).

Albedo Map. The color palette for each rendered shoe comes from the color distribution of a real shoe-tread photograph. Shoes tend to have only a handful of different colors across the entire tread. We identify the primary colors on real shoe-treads using the mean-shift algorithm [17]. Albedo maps for the rendered shoes are composed of these colors. First, we use depth maps to identify shoe-tread elements and segment out areas of the shoe that can have different colors. Then we assign colors to those segments from the color palette of a real shoe in the percentages in which they are present. Fig. 3 shows one example.

Light environment. Online retail stores use specialized diffuse lighting rigs to capture shoe photos. We create a similar lighting environment for our rendered images. Shoes are photographed with bright diffuse white light from all directions and some optional directional light. We use a total of 17 different light configurations. One light configuration is simply diffuse light coming from all directions. Eight light configurations consist of single light bulbs shining from eight directions around the shoe in addition to the diffuse white light. The remaining eight are similar but contains two light bulbs at 120° to each other. The supplement has further details.

4.2. Online Shoe Treads for Training and Prediction

Online retailers [1, 2] adopt photos of shoes for advertisement, which include shoe-tread images. Real-train (3,543), cf. Table 1, consists of such shoe-tread images and masks computed by a simple network to segment out the shoe-treads. This dataset does not contain any ground-truth and consists only of new, athletic shoes.

4.3. Lab Data for Validation

Real-val. To quantitatively benchmark methods, we collect paired shoe-tread images and ground-truth prints in a

Figure 4: Generating pseudo albedo maps from shoe-tread images. We show two pairs. We run the mean-shift algorithm [17] on a shoe-tread image to group RGB pixels, resulting in the corresponding pseudo albedo map. We use the pseudo albedo maps as supervision signals to train the decomposer (cf. Fig. 1).

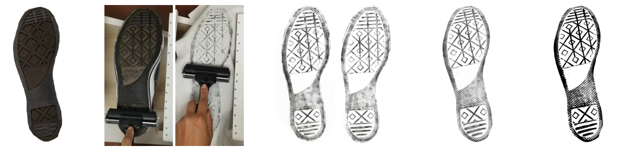
lab environment. Fig. 5 summarizes the procedure. We photograph shoes by placing them inside a light box with a ring light on top. We collect prints from those shoes by painting the treads with a thin layer of relief ink and pressing absorbent white papers onto the shoe-treads. This method of collecting shoeprints is called the *block printing technique* and is one of several techniques used in the forensics community to collect reference footwear impressions [9]. To improve print quality, we collect 2-3 prints for each shoe and average them after alignment to the shoe-tread. We use thin-plate splines [13] with a smoothness parameter of 0.5 for alignment. We threshold the average print as the final ground-truth shoeprint. Real-val contains 22 new-athletic shoes, 6 new formal shoes, and 8 used athletic shoes. The formal and used shoes are not present during training and thus serve as novel examples in evaluation.

Real-FID-val. We introduce the second validation set consisting of shoeprints from the FID300 dataset [32] and shoe-tread images separately downloaded from online retailers (i.e., these images are disjoint from those in the real-train set). We find matched FID300 prints (used as the ground-truth) and the downloaded shoe-tread images, and align them manually. Real-FID-val contains 41 new, athletic shoe-tread images with corresponding ground-truth shoeprints and masks to segment out the shoe-treads.

5. Methodology

We now introduce our *ShoeRensics*, a pipeline that trains a depth predictor for real images I_R by incorporating unsupervised adversarial domain adaptation and intrinsic image decomposition techniques. Given synthetic images I_S with their corresponding ground-truth intrinsics (albedo X^a_S, depth X^d_S, normal X^n_S, and light X^l_S) and unlabeled real images I_R, our goal is to train a model to predict depth d_R for real images I_R. Fig. 1 overviews our training pipeline. The main components of our pipeline are a translator $\mathcal{G}_{S\rightarrow R}$ to stylize synthetic images as real images, a decomposer $\mathcal{F}$ for intrinsic image decomposition, and a renderer $\mathcal{R}$ to reconstruct the input images from their intrinsic components.

Synthetic-only Training. We train a decomposer $\mathcal{F}$ and the renderer $\mathcal{R}$ in a supervised manner on syn-train. For an input image, the decomposer predicts depth $\hat{X}^d_S$, albedo $\hat{X}^a_S$, normal $\hat{X}^n_S$, and light $\hat{X}^l_S$. The renderer $\mathcal{R}$ learns to



(a) shoe-tread image (b) ink shoe-tread and collect print (c) 2 prints for the same shoe-tread (d) align to shoe-tread and average prints (e) final binary ground-truth

Figure 5: We collect a validation set of ground-truth shoeprints from shoes in a lab environment. (a) shows an example shoe. (b) It is painted with a thin layer of relief ink, and a paper sheet is pressed evenly onto the shoe-tread using a roller. (c) We repeat this to get 2-3 different prints. (d) We align these prints to the shoe-tread using thin-plate spline [13] and (e) threshold their average to obtain the final ground-truth shoeprint, which has better coverage.

reconstruct the input image from these predicted intrinsic components. To train the decomposer $\mathcal{F}$, we use an $\mathcal{L}_1$ loss to learn for depth, albedo, and normal prediction and a cross-entropy loss $\mathcal{L}_{CE}$ to learn for light (treating light prediction as a K-way classification problem given the limited light sources). We minimize the overall loss below:

$$\mathcal{L}_{sup} = \lambda_l \mathcal{L}_{CE}(\hat{X}_S^l, X_S^l) + \sum_{\kappa \in \{d,a,n\}} \lambda_\kappa \mathcal{L}_1(\hat{X}_S^\kappa, X_S^\kappa). \quad (1)$$

where λ's are hyperparamters controlling loss terms for the intrinsic components. To learn the renderer $\mathcal{R}$, we simply minimize the $\mathcal{L}_1$ loss between the original and rendered images, i.e., $\mathcal{L}_1(I_S, \mathcal{R}(X_S^d, X_S^a, X_S^n, X_S^l))$. Note that depth prediction is our main focus, and we find learning with decomposer and renderer significantly helps depth learning (cf. Fig. 1, Table 2). A model trained on synthetic data only does not work effectively well on real data due to the notorious synthetic-real domain gap. We address this issue using the techniques below.

Mitigating domain gap by image translation. Previous work [25, 70] addresses the domain gap between image sources by translating images from one domain to the other. We adopt a similar approach and translate our synthetic images to realistic ones by training a translator $\mathcal{G}_{S\rightarrow R}$. We train another $\mathcal{G}_{R\rightarrow S}$ that translates real images to synthetic style. Discriminators $\mathcal{D}_R(I)$ and $\mathcal{D}_S(I)$ are learned simultaneously to discriminate translated images and used for training translators. This is known as the adversarial domain adaptation [25]. We further translate the translated synthetic/real images back to the original domain and use a cycle loss between the resulting and the initial images to ensure that structure and content are preserved during translation. The following losses train the translators [25, 70]:

$$\begin{aligned}
&\mathcal{L}_{GAN}^{S\rightarrow R}(I_R, I_S) = \log \mathcal{D}_R(I_R) + \log(1 - \mathcal{D}_R(\mathcal{G}_{S\rightarrow R}(I_S)))\\
&\mathcal{L}_{GAN}^{R\rightarrow S}(I_S, I_R) = \log \mathcal{D}_S(I_S) + \log(1 - \mathcal{D}_S(\mathcal{G}_{R\rightarrow S}(I_R)))\\
&\mathcal{L}_{tran} = \mathcal{L}_{GAN}^{S\rightarrow R}(I_R, I_S) + \mathcal{L}_{GAN}^{R\rightarrow S}(I_S, I_R)\\
&\mathcal{L}_{cyc} = \mathcal{L}_1(\mathcal{G}_{R\rightarrow S}(\mathcal{G}_{S\rightarrow R}(I_S)), I_S) +\\
&\qquad\quad \mathcal{L}_1(\mathcal{G}_{S\rightarrow R}(\mathcal{G}_{R\rightarrow S}(I_R)), I_R) \qquad (2)
\end{aligned}$$

With $\mathcal{G}_{S\rightarrow R}(I_S)$, we translate syn-train images and keep their corresponding ground-truth intrinsics unchanged. We use such translated data to finetune the renderer $\mathcal{R}$.

Mitigating domain gap by image reconstruction. We additionally use an image reconstruction loss to address the domain gap [28]. We reconstruct a real image from its decomposed intrinsic components using the trained renderer $\mathcal{R}$, which we freeze after finetuning on translated synthetic data. We use $\mathcal{R}$ to regularize the training of the decomposer $\mathcal{D}$ on real images. Denoting reconstructed real image as $\hat{I}_R := \mathcal{R}(X_R^d, X_R^a, X_R^n, X_R^l)$, we minimize the difference between the original image I_R and its reconstruction $\hat{I}_R$ using an $\mathcal{L}_1$ loss, i.e., $\mathcal{L}_1(\hat{I}_R, I_R)$.

Mitigating domain gap by feature alignment. We further adopt the feature alignment technique to mitigate the domain gap [70]. Specifically, we learn an adversarial discriminator $\mathcal{D}_{feat}$ to discriminate *features* extracted by the decomposer for the real images and the translated synthetic images. We use this as a loss in training the decomposer and update the discriminator $\mathcal{D}_{feat}$ while training of the decomposer. This encourages the decomposer to extract features on real data that are indistinguishable from synthetic data, thus helping mitigate the domain gap.

Exploiting Pseudo Albedo. Shoe-treads, like many other man-made objects such as cars and other toys, tend to have piece-wise constant albedo. Building on this observation, we create pseudo albedo for the real data by grouping pixels with the mean-shift algorithm [17]. Fig. 4 shows an example pseudo albedo on two real shoes. As pseudo albedo is not ideal as ground-truth, we use it to learn an albedo predictor through the the decomposer. We find this produces better albedo maps than the pseudo ground-truth (see analysis in the supplement). To learn albedo prediction, we minimize the $\mathcal{L}_1$ loss, i.e., $\mathcal{L}_1(\hat{X}_R^a, \text{MS}(I_R))$, where MS is the mean-shift clustering algorithm.

Stage-wise Training is common in training multiple modules, particularly with GAN discriminators. Our training paradigm contains four stages. First, we train the decomposer $\mathcal{F}$ and renderer $\mathcal{R}$ on syn-train. Second, we train the image translators and discriminators $G_{S\rightarrow R}$, $\mathcal{G}_{R\rightarrow S}$, $\mathcal{D}_R$, and $\mathcal{D}_S$ with Eq. 2. Third, we finetune $\mathcal{R}$ using the translated synthetic images by $\mathcal{G}_{S\rightarrow R}$. Finally, we freeze $\mathcal{R}$ and $\mathcal{G}_{S\rightarrow R}$ and finetune $\mathcal{F}$ on translated synthetic images and real images using losses described above.

6. Experiments

We validate our *ShoeRinsics* and compare it against prior methods of depth prediction on our benchmark. We start with implementation details, followed by a visual comparison and quantitative evaluation, and conduct an ablation study and analysis of why *ShoeRinsics* outperforms the prior art.

6.1. Implementation

Training specifics. Instead of using high-resolution images (405x765) from the training set, we crop patches

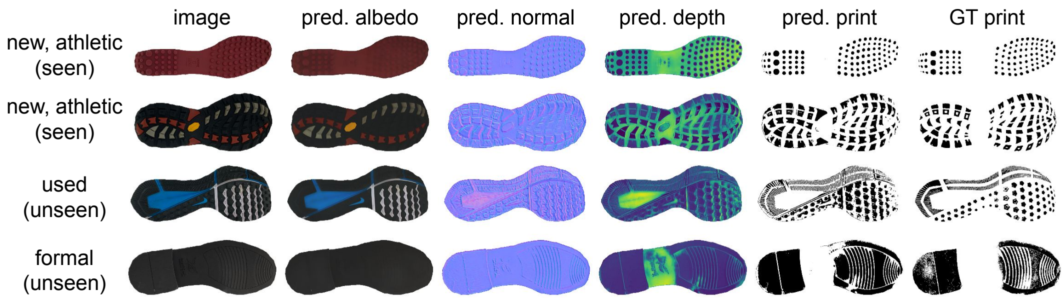


Figure 6: On images of the real-val set, we visualize *ShoeRinsics*'s predictions including depth thresholding which generates predicted prints. Our method *ShoeRinsics* produces visually appealing intrinsic decompositions (depth, albedo, and normal). Importantly, on novel shoe-tread displayed in the last bottom rows, *ShoeRinsics* produces very good depth and shoeprints by comparing against the ground-truth shoeprints. To display the predicted prints, we threshold the predicted depth to best match the ground-truth print (Sec 3.2).

(128x128) to train the models. We find this yields better performance, as shown in the ablation study (Sec. 6.4). For a fair comparison, we train all models with patches for the same number of optimization steps. During training, we sample patches from random positions. We use Adam optimizer and set the learning rate as 1e-3 and 1e-4 for training the initial models (e.g., $\mathcal{F}$ and $\mathcal{R}$) and finetuning them, respectively. We set the batch size as 8 throughout our experiments. Recall that we train our model in stages (Sec. 5). We train for 20M iterations in the first two stages and 100K iterations in the last two stages.

Architectures. Our decomposer $\mathcal{F}$ and renderer $\mathcal{R}$ have a classic encoder-decoder structure as used in [28]. We modify the light prediction decoder to be a 17-way classifier (given that our synthetic data has only 17 lighting configurations). We also add residual connections between layers to predict full-resolution maps for intrinsic components (depth, albedo, and normal). Our translators and discriminators ($\mathcal{G}_{S\rightarrow R}, \mathcal{G}_{R\rightarrow S}, \mathcal{D}_R, \mathcal{D}_S$, and $\mathcal{D}_{feat}$) have the same structure as used in [25]. The $\mathcal{D}_{feat}$ is a convolutional network that uses a kernel size 3 to process the albedo, depth, and normal features. It further takes as input the features of the lighting prediction branch. That said, $\mathcal{D}_{feat}$ learns to discriminate features of all the intrinsic components.

Hyperparameter setting. We denote the combined hyperparameters as $\hat{\lambda} = (\lambda_a, \lambda_d, \lambda_n, \lambda_l)$ in Eq. 1. The decomposer $\mathcal{F}$ is trained with $\hat{\lambda} = (1, 1, 1, 0.1)$ in the first stage, and finetuned with $\hat{\lambda} = (1, 2, 1, 0.1)$ in the final stage. When finetuning, we set the weight to 3 for the reconstruction loss, 2 for the pseudo albedo loss, and 1 for the feature alignment. We set the hyperparameters via validation.

Test-time augmentation. During testing, we consider test-time augmentation [14, 24]. For each image, we produce 23 variants: 3 flips (horizontal, vertical, and vertical+horizontal), 4 rotations (angles $+5°$, $+10°$, $-5°$, and $-10°$), 4 scalings (scale factor 0.5, 0.8, 1.5, and 1.8), and 12 flip+rotation versions (three flips times four rotations). For each variant, we predict the depth and then transform back to the original coordinate frame. We average all the 24 depth maps as the final prediction.

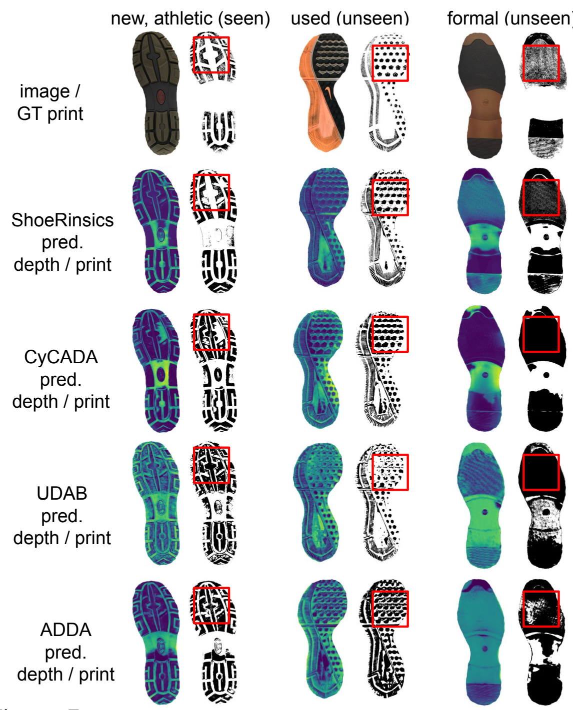


Figure 7: Comparison with the state-of-the-art methods of domain adaptation tailored to depth prediction on our real-val benchmark. Our *ShoeRinsics* performs better than others for both seen and unseen shoe categories as highlighted by the red boxes.

6.2. Qualitative Results of ShoeRinsics

We visualize predictions on the real-val images by our method *ShoeRinsics* in Fig. 6. *ShoeRinsics* predicts good depth maps, the thresholding of which generates shoeprints that match the ground-truth prints. As a byproduct, our method also makes visually appealing predictions on other intrinsic components. We compare our predictions with those made by other methods on real-val (Fig. 7) and real-FID-val (Fig. 8). Clearly, our *ShoeRinsics* produces more reasonable visuals (depth and shoeprints) than the compared methods. The supplement has further visualizations.

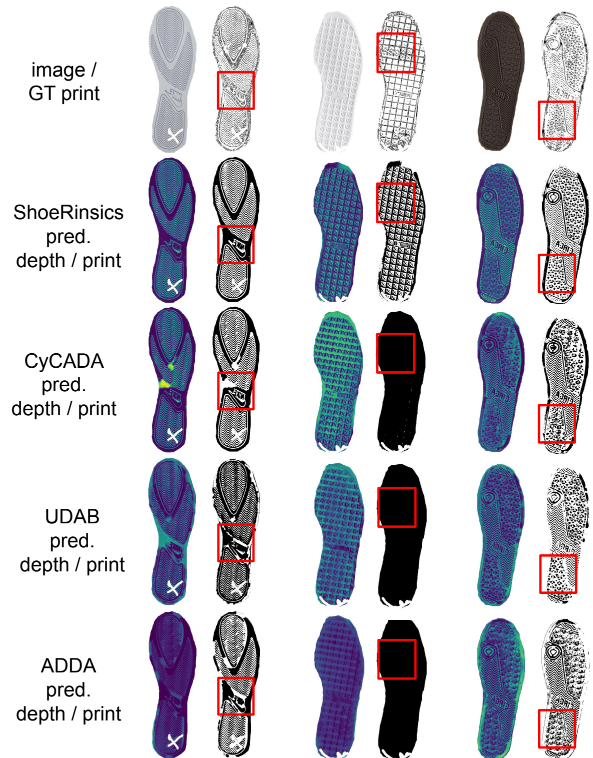


Figure 8: Comparison with the state-of-the-art methods for depth prediction and domain adaptation [19, 25, 55] on real-FID-val. Clearly, our *ShoeRinsics* produces shoeprints which are visually closer to the ground-truth than previous methods.

6.3. State-of-the-art Comparison

ShoeRinsics outperforms prior methods in most of the validation examples (details in the supplement). Table 2 and 3 list comparisons as analyzed below.

Comparison with intrinsic image decomposition. We compare our *ShoeRinsics* and RIN [28], which learns for intrinsic image decomposition. As RIN [28] emphasizes normal prediction to represent shapes, we use the standard Frankot-Chellappa algorithm [15] to integrate the normals towards depth maps. Compared to [28], our *ShoeRinsics* explicitly incorporates domain adaptation in the image and feature space. Doing so helps mitigate the synthetic-real domain gap. As a result, *ShoeRinsics* outperforms RIN on both real-val and real-FID-val (Table 2 and 3). On real-val, it performs better than RIN by 20.5% mIoU on the (*seen new-athletic*) shoes, by 8.1% mIoU on the *formal unseen* shoes, by 11.4% mIoU on *used unseen* shoes. On real-FID-val, *ShoeRinsics* improves IoU by 5.6% mIoU over RIN.

Comparison with domain adaptation. Table 2 and 3 clearly show that our *ShoeRinsics* consistently outperforms the compared domain adaptation methods (ADDA [55], UDAB [19], and CyCADA [25]) on both the real-val and real-FID-val datasets. From ablation studies, as shown in the lower panel of Table 2, we see that using the renderer (cf. Fig. 1) and the decomposer (that learns to predict albedo, normal, and lighting as auxiliary supervisions) greatly improves the performance. Qualitative comparison on real-val in Fig. 7 and real-FID-val in Fig. 8 show that depth maps and the corresponding prints predicted by our *ShoeRinsics* have richer textures and better-aligned patterns to the RGB input. When exploiting test-time augmentation (cf. *ShoeRinsics* w/ test-time aug), we boost the performance from mIoU = 46.8% to 49.0% on real-val and from mIoU=31.6% to 32.0% on real-FID-val.

Table 2: Benchmarking on real-val. We use IoU as the metric (in %), and break down the analysis for different shoe categories (*new-athletic* shoes seen during training, and *formal* and *used* shoes unseen in training). We compute mean IoU (mIoU) over all validation examples. Training on only synthetic data yields poor performance, whereas our *ShoeRinsics* performs the best on both seen and unseen categories. This clearly demonstrates the benefit of combining synthetic-to-real domain adaptation with intrinsic decomposition. The ablation study (bottom panel) shows that each individual component (discriminator, translator, and renderer, cf. Fig. 1) helps improve shoeprint prediction. Lastly, from our syn-only ablation, decomposing to all intrinsic components performs better than training a depth predictor for shoeprint prediction, further demonstrating that incorporating intrinsic decomposition helps close synthetic-to-real domain gaps. Exploiting test-time augmentation boosts performance from *mIoU*=46.8 to 49.0.

Method	*New-Athletic (seen)*	*Formal (unseen)*	*Used* (unseen)	*mIoU*
RIN [28]	30.0	39.7	24.4	30.4
ADDA [55]	46.5	41.4	27.2	41.4
UDAB [19]	46.0	40.4	29.6	41.4
CyCADA [25]	48.8	43.9	34.5	44.8
syn-only, depth only	41.3	41.2	28.4	38.4
syn-only, all intrinsics	41.8	41.5	27.1	38.5
ShoeRinsics	50.5	47.8	35.8	46.8
w/o discriminator	48.2	39.9	33.6	43.6
w/o translator	49.0	42.8	31.4	44.0
w/o renderer	49.0	46.4	34.7	45.4
ShoeRinsics w/ aug	**52.4**	**52.9**	**36.9**	**49.0**

Table 3: Benchmarking on real-FID-val. We report mean IoU (mIoU) over validation examples. *ShoeRinsics* outperforms previous methods and improves further with test-time augmentation.

	RIN [28]	ADDA [55]	UDAB [19]	CyCADA [25]	**ShoeRinsics**	**ShoeRinsics** w/ aug
mIoU	26.0	27.2	29.0	31.2	31.6	**32.0**

Performance on real-val vs real-FID-val. All the methods show lower mIoU numbers on real-FID-val compared to real-val. This is owing to the noisy ground-truth prints of real-FID-val (see Fig. 9). Note that the FID prints are obtained by pressing gelatin lifters onto dusty shoe-treads followed by scanning the lifters [32]. This means that the shoeprints can be noisy as the contact surfaces do not leave a full print. In contrast, for real-val shoeprints,

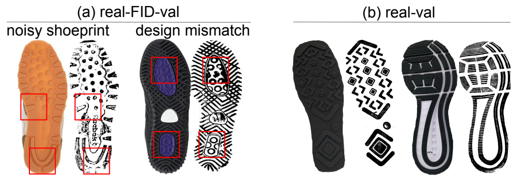


Figure 9: Comparison between real-FID-val (a) and real-val (b). The shoeprints from real-FID-val are noisy and slightly misaligned with the corresponding shoe-treads. In contrast, shoeprints of real-val contain the entire contact surfaces and are well aligned with the corresponding shoe-tread images.

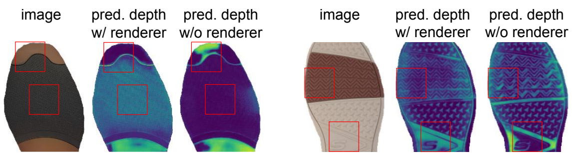


Figure 10: Training *ShoeRinsics* with the renderer (which allows using the reconstruction loss) produces visibly better depth than without. Using the renderer encourages the decomposer to output depth maps that contain fine-grained details because it penalizes coarse predictions through the image reconstruction loss. That said, the renderer regularizes the learning for depth prediction by exploiting auxiliary supervisions from other intrinsic components (albedo, normal, and lighting).

we minimize such noise and get more even coverage by averaging over multiple prints for the same shoe. Moreover, while real-val consists of image and print pairs of the exact same shoe, real-FID-val consists of prints from [32] with our manually discovered shoe-tread images, meaning that they might not be well aligned, as visually seen in Fig. 9.

6.4. Ablation Study

We conduct an ablation study (cf. Table 2 bottom panel) on the modules in *ShoeRinsics*, including feature alignment (by learning discriminator $\mathcal{D}_{feat}$ in the feature space), translator $\mathcal{G}_{S\rightarrow R}$, and renderer $\mathcal{R}$. All three modules aim to mitigate synthetic-real domain gaps. We also study whether predicting intrinsic components (albedo, normal, and lighting) helps depth prediction and whether patch-based learning is better than full-image learning.

Effect of feature alignment by discriminator $\mathcal{D}_{feat}$. *ShoeRinsics w/o discriminator* removes the feature discriminator $\mathcal{D}_{feat}$ but keeps all the other modules. It yields 43.6% mIoU, 3.2% mIoU lower than *ShoeRinsics* (cf. Table 2). This demonstrates the effectiveness of $\mathcal{D}_{feat}$ for mitigating domain gaps by aligning features.

Effect of image translator $\mathcal{G}_{S\rightarrow R}$. *ShoeRinsics w/o translator* drops the translators but keeps other components, achieving 44.0% mIoU, 2.8% mIoU lower than *ShoeRinsics* (cf. Table 2). This shows the effectiveness of using translators to close the synthetic-real domain gap.

Effect of the reconstruction loss by the renderer $\mathcal{R}$. *ShoeRinsics w/o renderer* drops the renderer from *ShoeRinsics*, leading to 45.4% mIoU, 1.4% mIoU lower than

Figure 11: Failure cases. *ShoeRinsics* performs poorly in the presence of complex materials (e.g., translucence).

ShoeRinsics (cf. Table 2). This validates the effectiveness of the renderer. Fig. 10 visualizes depth predictions with and without the renderer during training. Clearly, with the renderer, the predicted depth has better high-frequency textures. See the caption of Fig. 10 for details.

All intrinsics vs depth only. Comparing "syn-only, depth only" and "syn-only, all intrinsics" in Table 2, we see that learning to predict all intrinsics performs slightly better (38.5% vs. 38.4%). Importantly, this allows using the renderer as the reconstruction loss to regularize the training on real images, yielding significantly better results in the final *ShoeRinsics* (46.8% mIoU).

Patches vs. full-resolution images. We compare the depth prediction performance by training the decomposer on patches versus full-resolution images of the synthetic data. We find that the former (patch-based) achieves 38.5% mIoU (cf. Table 2) as opposed to 36.5% mIoU for the latter (not shown in the table). This demonstrates the benefit of depth learning on patches over whole images in this setup.

6.5. Failure Cases

We analyze failure cases of *ShoeRinsics* in Fig. 11. We find that our method performs poorly on shoes with complex materials. One reason is that the syn-train data does not contain any complex materials. Future work may explore richer synthetic datasets to improve performance.

7. Conclusion

Motivated by constructing a database of shoeprints for forensic use, we introduce a problem of predicting depth for shoe-tread photos collected by online retailers. Because these photos do not have ground-truth depth, we exploit synthetic images (containing shoe-treads and ground-truth intrinsics including depth, albedo, normal, and lighting). We study domain adaptation and intrinsic image decomposition techniques and propose a method termed *ShoeRinsics* to train for depth prediction. Our experiments demonstrate consistent improvements of *ShoeRinsics* over previous methods on this task. We expect future algorithmic explorations on this task from the perspective of domain adaptation, depth prediction, and intrinsic decomposition.

Acknowledgements. This work was funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreements, 70NANB15H176 and 70NANB20H019.

References

[1] 6pm. http://www.6pm.com.

[2] Zappos. http://www.zappos.com.

[3] Gharsa AlGarni and Madina Hamiane. A novel technique for automatic shoeprint image retrieval. *Forensic science international*, 181(1-3):10–14, 2008.

[4] Hassan Abu Alhaija, Siva Karthik Mustikovela, Justus Thies, Varun Jampani, Matthias Nießner, Andreas Geiger, and Carsten Rother. Intrinsic autoencoders for joint neural rendering and intrinsic image decomposition, 2021.

[5] Sayyad Alizadeh and Cemal Kose. Automatic retrieval of shoeprint images using blocked sparse representation. *Forensic science international*, 277:103–114, 2017.

[6] Somaya Almaadeed, Ahmed Bouridane, Danny Crookes, and Omar Nibouche. Partial shoeprint retrieval using multiple point-of-interest detectors and sift descriptors. *Integrated Computer-Aided Engineering*, 22(1):41–58, 2015.

[7] Riza Alp Guler, George Trigeorgis, Epameinondas Antonakos, Patrick Snape, Stefanos Zafeiriou, and Iasonas Kokkinos. Densereg: Fully convolutional dense shape regression in-the-wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6799–6808, 2017.

[8] Mohammad Haris Baig and Lorenzo Torresani. Coupled depth learning. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10. IEEE, 2016.

[9] William J Bodziak. *Footwear impression evidence: detection, recovery, and examination*. CRC Press, 2017.

[10] Ahmed Bouridane, A Alexander, Mokhtar Nibouche, and Danny Crookes. Application of fractals to the detection and classification of shoeprints. In *Proceedings 2000 International Conference on Image Processing (Cat. No. 00CH37101)*, volume 1, pages 474–477. IEEE, 2000.

[11] Robin Bowen and Jessica Schneider. Forensic databases: paint, shoe prints, and beyond. *NIJ Journal*, 258:34–38, 2007.

[12] Philip De Chazal, John Flynn, and Richard B Reilly. Automated processing of shoeprint images based on the fourier transform for use in forensic science. *IEEE transactions on pattern analysis and machine intelligence*, 27(3):341–350, 2005.

[13] Jean Duchon. Splines minimizing rotation-invariant seminorms in sobolev spaces. In *Constructive theory of functions of several variables*, pages 85–100. Springer, 1977.

[14] David A. Forsyth and Jason J. Rock. Intrinsic image decomposition using paradigms. *CoRR*, abs/2011.10512, 2020.

[15] Robert T. Frankot and Rama Chellappa. A method for enforcing integrability in shape from shading algorithms. *IEEE Transactions on pattern analysis and machine intelligence*, 10(4):439–451, 1988.

[16] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2002–2011, 2018.

[17] Keinosuke Fukunaga and Larry Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on information theory*, 21(1):32–40, 1975.

[18] Ryo Furukawa, Ryusuke Sagawa, and Hiroshi Kawasaki. Depth estimation using structured light flow–analysis of projected pattern flow on an object's surface. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4640–4648, 2017.

[19] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France, 07–09 Jul 2015. PMLR.

[20] Ravi Garg, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European conference on computer vision*, pages 740–756. Springer, 2016.

[21] Mourad Gueham, Ahmed Bouridane, and Danny Crookes. Automatic recognition of partial shoeprints based on phase-only correlation. In *2007 IEEE International Conference on Image Processing*, volume 4, pages IV–441. IEEE, 2007.

[22] Mourad Gueham, Ahmed Bouridane, and Danny Crookes. Automatic classification of partial shoeprints using advanced correlation filters for use in forensic science. In *2008 19th International Conference on Pattern Recognition*, pages 1–4. IEEE, 2008.

[23] Mourad Gueham, Ahmed Bouridane, Danny Crookes, and Omar Nibouche. Automatic recognition of shoeprints using fourier-mellin transform. In *2008 NASA/ESA Conference on Adaptive Hardware and Systems*, pages 487–491. IEEE, 2008.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[25] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. CyCADA: Cycle-consistent adversarial domain adaptation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1989–1998. PMLR, 10–15 Jul 2018.

[26] Derek Hoiem, Alexei A Efros, and Martial Hebert. Automatic photo pop-up. In *ACM SIGGRAPH 2005 Papers*, pages 577–584. 2005.

[27] Wenzel Jakob. Mitsuba renderer, 2010. http://www.mitsuba-renderer.org.

[28] Michael Janner, Jiajun Wu, Tejas D. Kulkarni, Ilker Yildirim, and Joshua B. Tenenbaum. Self-supervised intrinsic image decomposition. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 5938–5948, Red Hook, NY, USA, 2017. Curran Associates Inc.

[29] Min-Quan Jing, Wei-Jong Ho, and Ling-Hwei Chen. A novel method for shoeprints recognition and classification. In *2009*

International conference on machine learning and cybernetics, volume 5, pages 2846–2851. IEEE, 2009.

[30] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

[31] Bailey Kong, James Supancic, Deva Ramanan, and Charless C Fowlkes. Cross-domain image matching with deep feature maps. *International Journal of Computer Vision*, 127(11):1738–1750, 2019.

[32] Adam Kortylewski, Thomas Albrecht, and Thomas Vetter. Unsupervised footwear impression analysis and retrieval from crime scene data. In *Asian conference on computer vision*, pages 644–658. Springer, 2014.

[33] Adam Kortylewski and Thomas Vetter. Probabilistic compositional active basis models for robust pattern recognition. In *BMVC*, 2016.

[34] Scott Krig. Interest point detector and feature descriptor survey. In *Computer vision metrics*, pages 187–246. Springer, 2016.

[35] Yevhen Kuznietsov, Jorg Stuckler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6647–6655, 2017.

[36] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016.

[37] Jun Li, Reinhard Klein, and Angela Yao. A two-streamed network for estimating fine-scaled depth maps from single rgb images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3372–3380, 2017.

[38] Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2475–2484, 2020.

[39] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2041–2050, 2018.

[40] Mingsheng Long, Guiguang Ding, Jianmin Wang, Jiaguang Sun, Yuchen Guo, and Philip S Yu. Transfer sparse coding for robust image representation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 407–414, 2013.

[41] Zhanyu Ma, Yifeng Ding, Shaoguo Wen, Jiyang Xie, Yifeng Jin, Zhongwei Si, and Haining Wang. Shoe-print image retrieval with multi-part weighted cnn. *IEEE Access*, 7:59728–59736, 2019.

[42] Thomas Nestmeyer, Jean-Franccois Lalonde, Iain Matthews, and Andreas Lehrmann. Learning physics-guided face relighting under directional light. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5124–5133, 2020.

[43] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[44] Rene Ranftl, Vibhav Vineet, Qifeng Chen, and Vladlen Koltun. Dense monocular depth estimation in complex dynamic scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4058–4066, 2016.

[45] Imad Rida, Lunke Fei, Hugo Proencca, Amine Nait-Ali, and Abdenour Hadid. Forensic shoe-print identification: a brief survey. *arXiv preprint arXiv:1901.01431*, 2019.

[46] Anirban Roy and Sinisa Todorovic. Monocular depth estimation using neural regression forest. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5506–5514, 2016.

[47] Ashutosh Saxena, Sung Chung, and Andrew Ng. Learning depth from single monocular images. *Advances in neural information processing systems*, 18, 2005.

[48] Ashutosh Saxena, Sung H Chung, and Andrew Y Ng. 3-d depth reconstruction from a single still image. *International journal of computer vision*, 76(1):53–69, 2008.

[49] Soumyadip Sengupta, Jinwei Gu, Kihwan Kim, Guilin Liu, David W. Jacobs, and Jan Kautz. Neural inverse rendering of an indoor scene from a single image. In *International Conference on Computer Vision (ICCV)*, 2019.

[50] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D. Castillo, and David W. Jacobs. Sfsnet: Learning shape, refectance and illuminance of faces in the wild. In *Computer Vision and Pattern Regognition (CVPR)*, 2018.

[51] Daeyun Shin, Zhile Ren, Erik B Sudderth, and Charless C Fowlkes. 3d scene reconstruction with multi-layer depth and epipolar transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2172–2182, 2019.

[52] Michael B. Smith. The forensic analysis of footwear impression evidence. *Forensic Science Communications*, 2009. https://archives.fbi.gov/archives/about-us/lab/forensic-science-communications/fsc/july2009/review.

[53] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016.

[54] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7472–7481, 2018.

[55] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2962–2971, 2017.

[56] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[57] Xinnian Wang, Chi Zhang, Yanjun Wu, and Yingying Shu. A manifold ranking based method using hybrid features for crime scene shoeprint retrieval. *Multimedia Tools and Applications*, 76(20):21629–21649, 2017.

[58] Zhibo Wang, Xin Yu, Ming Lu, Quan Wang, Chen Qian, and Feng Xu. Single image portrait relighting via explicit multiple reflectance channel modeling. *ACM Trans. Graph.*, 39(6), Nov. 2020.

[59] Shangzhe Wu, Ameesh Makadia, Jiajun Wu, Noah Snavely, Richard Tucker, and Angjoo Kanazawa. De-rendering the world's revolutionary artefacts. In *CVPR*, 2021.

[60] Junyuan Xie, Ross Girshick, and Ali Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *European conference on computer vision*, pages 842–857. Springer, 2016.

[61] Yoram Yekutieli, Yaron Shor, Sarena Wiesner, and Tsadok Tsach. Expert assisting computerized system for evaluating the degree of certainty in 2d shoeprints. *The US Department of Justice: Washington, DC, USA*, 2012.

[62] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5684–5693, 2019.

[63] Si Li Yongjie Zhu, Yinda Zhang and Boxin Shi. Spatially-varying outdoor lighting estimation from intrinsics. In *CVPR*, 2021.

[64] Ye Yu, Abhimetra Meka, Mohamed Elgharib, Hans-Peter Seidel, Christian Theobalt, and Will Smith. Self-supervised outdoor scene relighting. In *European Conference on Computer Vision (ECCV)*, 2020.

[65] Yang Zhang, Huanzhang Fu, Emmanuel Dellandréa, and Liming Chen. Adapting convolutional neural networks on the shoeprint retrieval for forensic use. In *Chinese Conference on Biometric Recognition*, pages 520–527. Springer, 2017.

[66] Yunhan Zhao, Shu Kong, and Charless Fowlkes. Camera pose matters: Improving depth prediction by mitigating pose distribution bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15759–15768, 2021.

[67] Yunhan Zhao, Shu Kong, Daeyun Shin, and Charless Fowlkes. Domain decluttering: Simplifying images to mitigate synthetic-real domain shift and improve depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3330–3340, 2020.

[68] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. T2net: Synthetic-to-realistic translation for solving single-image depth estimation tasks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018.

[69] Hao Zhou, Xiang Yu, and David Jacobs. Glosh: Global-local spherical harmonics for intrinsic image decomposition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7819–7828, 2019.

[70] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.