# Teaching Cameras to Feel: Estimating Tactile Physical Properties of Surfaces From Images

Matthew Purri and Kristin Dana

Rutgers University

**Abstract.** The connection between visual input and tactile sensing is critical for object manipulation tasks such as grasping and pushing. In this work, we introduce the challenging task of estimating a set of tactile physical properties from visual information. We aim to build a model that learns the complex mapping between visual information and tactile physical properties. We construct a first of its kind image-tactile dataset with over 400 multiview image sequences and the corresponding tactile properties. A total of fifteen tactile physical properties across categories including friction, compliance, adhesion, texture, and thermal conductance are measured and then estimated by our models. We develop a cross-modal framework comprised of an adversarial objective and a novel visuo-tactile joint classification loss. Additionally, we introduce a neural architecture search framework capable of selecting optimal combinations of viewing angles for estimating a given physical property.

**Keywords:** Cross-Modal, Visuo-Tactile, Viewpoint Selection, Physical Property Estimation, Neural Architecture Search, Tactile

## 1 Introduction

In real-world tasks such as grasp planning and object manipulation, humans infer physical properties of objects from visual appearance. Inference of surface properties is distinct from object recognition. For example in Figure 1a, the objects have different geometric shape; however, they share similar tactile physical properties. We can imagine what it would feel like to pick one up and handle it. Recognition can provide the semantic labels of the utensils, but tactile inference can provide the physical properties of the stainless steel. In this work, we introduce a computational model that learns the complex relationship between visual perception and the direct tactile physical properties of surfaces such as compliance, roughness, friction, stiction, and adhesive tack.

There are many instances where an accurate estimate of a surface's tactile properties is beneficial for automated systems. In colder climates for example, thin layers of ice form over driving surfaces dramatically decreasing the sliding friction of a road. Modern vision systems trained on autonomous driving datasets such as KITTI [18] or Cityscapes [10] can readily identify "road" pixels, but would not provide the coefficient of friction required for braking control. Another example is manufacturing garments or shoes that require precise manipulation of multiple types of materials. Delicate and smooth fabrics such as silk require

different handling than durable denim fabric. Also, robotic grasping and pushing of objects in a warehouse can benefit from surface property estimation to improve manipulation robustness.

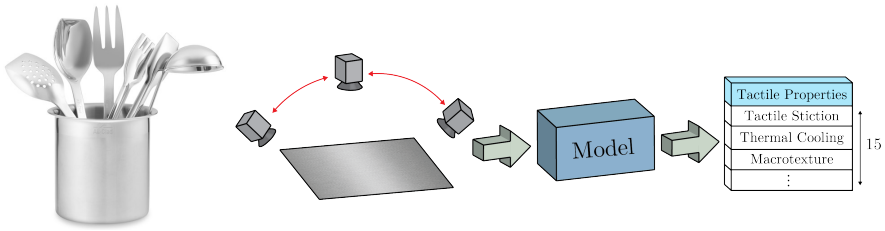


Fig. 1: **Material example and inference framework. Left:** An example of objects with different geometry and semantic labels that share common physical properties. **Right:** The proposed inference model receives images taken at various viewing angles and predicts a set of tactile property values.

In recent years, there has been increased interest in estimating physical properties of objects and forming learned physics engine models to infer the reaction of an object to controlled stimuli. Many of these methods passively observe interacting objects or select actions to stimulate object movement, learning to map visual information to physical properties by estimating the effect of the action on the scene [59, 58, 61, 23]. In these methods, geometric and physical properties of the objects are encoded into latent space, and a learned physics engine predicts the future state of the objects. The indirect representation of the object properties confounds high-level actions, such as pushing with the precise amount of force to overcome friction or ordering surfaces based on their roughness. In our work we estimate the physical properties directly, allowing attributes of objects to be utilized directly.

We formulate the challenge of estimating a surface's physical properties from visual information as a cross-modal translation problem. Cross-modal problems remain one of the frontiers in vision research. For example, many recent works relate disparate information streams such as videos with audio [40, 42, 1, 41] and images with captions [38, 28, 31]. Here, the problem of translating images into tactile properties provides a unique and challenging task. The visual and tactile modalities are not aligned in the same manner as audio and video and the scale discrepancy between images and tactile property vectors is vast.

To address this challenge, we create a dataset of 400+ surface image sequences and tactile property measurements. Given that humans have the ability to estimate tactile properties based on experience and visual cues, we expect autonomous systems to also be able to learn this complex mapping. While estimating surface properties, people often unconsciously move their heads, acquiring multiple views of a surface. Inspired by this, the captured images sequences comprise multiple viewing angles for each material surface.

Some challenges can be solved effectively using a single view of a scene, such as object classification, semantic segmentation, or image denoising. Whereas

tasks such as 3D geometry reconstruction and action recognition require more information than what a single image can typically provide. A rich literature of material recognition, a similar challenge to surface property estimation, has found advantages to using multiple viewpoints or illumination conditions to identify the material class of a surface. For example, reflectance disks [65, 66], optimal BRDF sampling [26, 34, 39], angular gradient images [62], 3D point clouds [12, 70], 4D light field images [56], and BRDF slices [55] all provide good material recognition performance with partial viewpoint or illumination sampling. These methods, however, rely on sampling at a fixed set of viewing angles. In this work, we allow our model to select the optimal partial BRDF for tactile property estimation, providing empirical insight for which viewpoints should be sampled in camera motion planning and physical property sensor designs.

The main objective of this work is to leverage the relation between the appearance of material surfaces and their tactile properties to create a network capable of mapping visual information to tactile physical properties. We have three main contributions. First, we develop a visual-tactile dataset named Surface Property Synesthesia dataset; with 400+ material surfaces imaged at 100 viewing angles and augmented with fifteen measured tactile physical properties (as listed in Table 1) measured by a BioTac Toccare tactile sensor. [1] Second, we propose a cross-modal learning framework with adversarial learning and cross-domain joint classification to estimate tactile physical properties from a single image. Third, we introduce two input information sampling frameworks that learn to select viewing angle combinations that optimize a given objective. Our results show that image-to-tactile estimation is challenging, but we have made a pioneering step toward direct tactile property estimation.

## 2 Related Work

**Cross-modal Learning and Translation** Cross-modal translation is defined as finding a function that maps information from one modality into its corresponding representation in a different modality. Prototypical approaches involve embedding data from different modalities into a learned space, from which generative functions map embeddings to their original representations. Recent works have applied this framework to various modality combinations including image-to-text [48, 68, 28, 45, 54], image-to-image [64, 69, 72], audio-to-image [52, 8], and audio-to-text [9]. Aytar et al. create a text, audio, and image multimodal retrieval system by leveraging large unaligned data sources [3]. Example-based translation methods such as retrieval systems rely on a dictionary (training data) when translating between modalities, whereas generative translation models directly produce translations. Generative translation [20, 33] is considered a more challenging problem than example-based translation because a generative function must be learned in addition to embedding functions. In [20], shared and

[1] Tactile measurements were done by SynTouch Inc., with the BioTac Toccare, and purchased by Rutgers.

domain-specific features are disentangled through auxiliary objectives to produce more realistic image-to-image translations. Generative translation models require large combinations of modality pairs to form a representative latent space. In this work, we introduce a new dataset containing a novel modality combination of image sequences and tactile physical properties.

**Visuo-Tactile** There is much interest in both the vision and robotics communities in giving robots the ability to understand the relation between visual and haptic information. A variety of challenges have been solved more efficiently with the addition of tactile information including object recognition [14], material classification [30], and haptic property estimation [17]. Calendra et al. combine a GelSight sensor [32, 64, 33, 6, 63] with an RGB camera to jointly predict grasp outcomes and plan action sequences for grasping [6]. Gao et al. improve the performance of a haptic adjective assignment task by fusing images and time sequence haptic measurements [17]. The aim of this work is not to improve the performance of a particular task but to find the relationship between visual information and touch, such that tactile properties can be estimated from visual information. Recently, works such as [64] and [33] similarly seek to learn a mapping between vision and touch either by learning a latent space for retrieval or by synthesizing realistic tactile signals from visual inputs. In contrast to these works, we directly generate tactile estimates and the representation of our tactile representation is a physical property vector instead of a tactile image. Most similar to our work, Zhang et al. estimate the coefficient of friction of a surface from a reflectance image [67]. We expand upon previous work to estimate a larger set of fifteen tactile properties including friction, texture, thermal conductance, compliance, and adhesion. Additionally, we assume that the visual information to our system consists of ordinary images obtained by standard RGB camera sensors.

**Viewpoint Selection** Given an oversampled set of images, how can we select the most useful subset of images from that set? In this work, we capture an oversampled sequence of images measured at consistent viewpoints via a gonioreflectometer. Inspired by viewpoint and illumination selection techniques for BRDFs [39, 26, 34, 62], our aim is to determine what combination of viewing angles produce the optimal output for physical property estimation. Nielsen et al. determine the minimum number of samples to reconstruct a measured BRDF by randomly sampling viewing and illumination angles and comparing their condition numbers [39]. Xue et al. capture pairs of images with small angular variations to generate improved angular gradients which serve as additional input for material recognition networks [62]. In contrast to viewpoint trajectory optimization [27, 60, 24], where the objective is to actively plan a viewing path in SO(3) space, to decrease task uncertainty, our objective is specifically to improve image-to-tactile estimation performance. In video understanding tasks, selectively sampling frames can enable efficient pattern analysis [71, 15, 49]. In [71], features from random subsets of frames of video are extracted and summed into a final representation to efficiently improve action recognition. Similarly, we seek

to sample the angular space of multiview images. However, our approach learns the optimal sampling for the task at hand.

Inspired by neural architecture search (NAS) approaches [44, 35, 73, 36], we learn to select a combination of viewing angles instead of choosing a handcrafted selection strategy similar to Xue et al. [62]. NAS methods are comprised of a search space, search strategy, and a performance estimation strategy. Generally, the search space of NAS is a set containing all possible layer functions and layer connections, whereas the search space for our problem are all combinations of viewing angles. To our knowledge, we are the first to utilize NAS for searching over the input space, instead of the overall architecture, resulting in a viewpoint selection. We propose two NAS frameworks for learning combinations of viewing angles which improve tactile physical property estimation as well as providing insight into what combinations of viewpoints are most informative for estimating a given physical property.

Table 1: **Tactile property acronyms.** Acronyms for each of the fifteen tactile properties measured by the Toccare device.

| | | | | | |
|---|---|---|---|---|---|
| fRS | Sliding Resistance | fST | Tactile Stiction | uCO | Microtexture Courseness |
| uRO | Microtexture Roughness | mRG | Macrotexture Regularity | mCO | Macrotexture Courseness |
| mTX | Macrotexture | tCO | Thermal Cooling | tPR | Thermal Persistance |
| cCM | Tactile Compliance | cDF | Local Deformation | cDP | Damping |
| cRX | Relaxation | cYD | Yielding | aTK | Adhesive Tack |

## 3 Surface Property Synesthesia Dataset

Synesthesia is the production of an experience relating to one sense by a stimulation of another sense. For example, when viewing an image of a hamburger you may unconsciously imagine the taste of the sandwich. In this work, images of surfaces are perceived and the tactile properties of that surface are estimated. To train a model for tactile physical property estimation, we collect a dataset named the *Surface Property Synesthesia Dataset* (SPS) consisting of pairs of RGB image sequences and tactile measurements. The dataset contains 400+ commonly found indoor material surfaces, including categories such as plastic, leather, wood, denim, and more as shown in Figure 2a. To our knowledge, this dataset contains the largest number of material surfaces of any visuo-tactile dataset, a necessity for learning the complex relation between vision and touch. A majority of the dataset belongs to four of the fifteen material categories. However, each category contains a diverse set of surfaces in terms of both color and pattern as shown in Figure 2b. *The dataset and source code are made publicly available.*[2]

**Tactile Data** The Biotac Toccare is a tactile sensing device that measures fifteen tactile physical properties of a surface and has been shown to identify

[2] https://github.com/matthewpurri/Teaching-Cameras-to-Feel

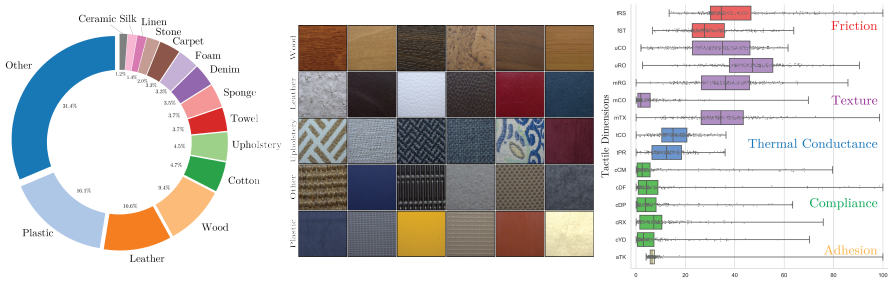


Fig. 2: **Dataset statistics. Left:** The distribution of material labels. The output of the models are tactile properties and not material labels. The other category includes materials with less than six occurrences. **Center:** A sampling of the 400+ materials in our dataset, which highlight the diversity in visual appearance. The other category includes materials such as metal, fur, corduroy, nylon, and more. **Right:** A box plot distribution for each tactile property. The acronym for each property is defined in Table 1.

materials more accurately than people [16]. The device is comprised of a BioTac sensor [51, 2, 29, 46] used to collect a series of time-varying signals and a staging system to consistently move the tactile sensor over a surface. Low-frequency fluid pressure, high-frequency fluid vibrations, core temperature change, and nineteen electrical impedances distributed along the sensor surface are recorded over time as the sensor is in contact with a surface. The signals are then converted into fifteen tactile physical properties whose values range from 0 to 100. Surface property measurements are gathered across several locations on the surface. Each tactile measurement is repeated five times.

The fifteen physical properties used to describe a surface, can be organized into five major tactile categories including friction, texture, thermal conductance, compliance, and adhesion as shown in Figure 2c. Specific descriptions for each of the fifteen properties are described in the supplementary material. The *texture* category represents both macro and micro-texture surface attributes which correspond to large and small height intensities along a surface. Both static and kinetic friction are included in the *friction* class. *Adhesion* describes the perceived effort to break contact with a surface with values semantically ranging from no adhesion to sticky. The rate of heat transferred from the BioTac sensor to the surface is described in the *thermal conductance* category. Surface deformation characteristics correspond to the *compliance* category.

**Vision Data** After tactile property measurements are obtained for each surface, images of the same surfaces are taken with a gonioreflectometer. The surfaces are imaged in a continuous manner from $-45°$ to $45°$ along the roll axis of the surface. For each material surface, 100 images are recorded. The yaw and pitch angles are constant throughout the imaging sequence. All images were taken under a mostly diffuse light source.

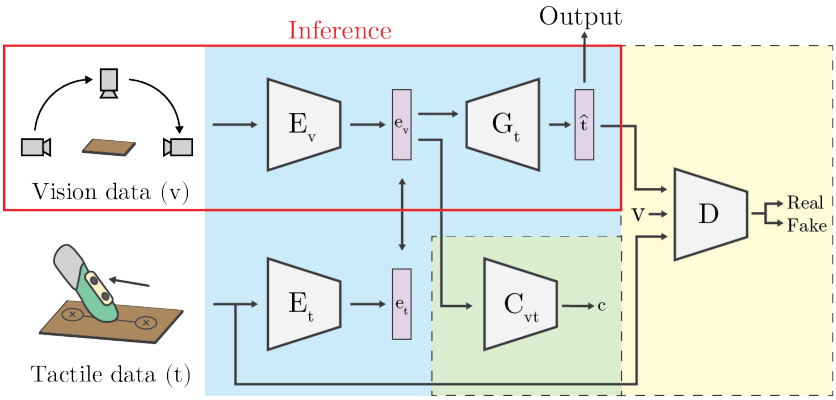


Fig. 3: **Overview of our proposed cross-modal framework.** The model is comprised of four modules: latent space encoding (blue), feature-classification (green), adversarial learning (yellow), and viewpoint selection (not displayed). The objective of this model is to generate precise tactile properties estimates $\hat{\mathrm{t}}_i$ given vision information $v_i$. Both visual and tactile information (measured with Toccare device) are embedded into a shared latent space through separate encoder networks and compared. A generator function $\mathrm{G_t}$ estimates tactile property values $\hat{\mathrm{t}}$ from the embedded visual vector $\mathrm{e_v}$. The discriminator network D learns to predict whether a tactile-visual pair is a real or synthetic example. An auxiliary classification network $\mathrm{C_{vt}}$ generates a visuo-tactile label given $\mathrm{e_v}$. The modules included in the red boundary represent the networks used during inference. Note, no tactile information is used during inference.

## 4 Methods

### 4.1 Mapping Vision to Touch

**Problem Definition** We model the problem of translation between two modalities as a cross-modal translation problem. We are specifically interested in the translation from images to tactile properties. Let $t_i \in \mathbb{R}^D$ represent tactile physical property vectors and $v_i \in \mathbb{R}^{3\times F\times H\times W}$ represent image sequences where $F$, $H$, and $W$ correspond to the number of frames, height, and width respectively. Instances of each modality are encoded through distinct encoding networks, $E_t$ and $E_v$, into a shared latent space. In this space, embedded visuo-tactile pairs $e_i^t$ and $e_i^v$ should be close and are encouraged to be near each other by a pairwise constraint $\mathcal{L}_{emb} = \|e_i^t - e_i^v\|_2^2$. Estimated tactile physical property vectors are created via a generative function $G_t$, given the embedded representation of the visual information as input $G_t(e_i^v) = \hat{t}_i$.

**Regression Baseline** To evaluate the capabilities of the cross-modal network, we compare its results to the results obtained from a regression network. The regression network encodes a single image of a material surface into a tactile physical property estimate omitting the intermediate latent representation and embedding constraint, $E_t(v_i) = \hat{t}_i$.

**Cross-Modal Network**

*Adversarial Objective* Inspired by multi-modal works [33, 69, 72], we augment our cross-modal framework with an adversarial objective in order to improve the quality of the tactile estimates. The input visual information is combined with the estimated tactile information and then used as input into a discriminator network to determine if the pair is real or fake. This process forms the following objective:

$$\begin{aligned}\mathcal{L}_{adv}(G_t, D) =& \mathbb{E}_{v,t}[\log D(v,t)] + \\ & \mathbb{E}_{v,t}[\log(1 - D(v, G_t(e_v))] + \mathbb{E}_{v,t}[\|G_t(e_v) - t\|_2],\end{aligned} \tag{1}$$

where the generator $G_t$ attempts to generate realistic tactile property vectors that are conditioned on the embedded visual information $e_v$ while the discriminator $D$ tries to distinguish between real versus fake visuo-tactile pairs. In prior work on conditional GANs [22, 11, 43], the input and output of the generator are the same dimension whereas the input of our generator can be a sequence of images and the output is a low dimensional vector. In order to handle the scale difference, we combine the tactile property estimation with the feature vector output of a single image instead of the full resolution image. The feature vector is generated via the image encoding network $E_v$.

*Classification Objective* A key to forming a latent space that is well conditioned for a given objective is to add constraints to that space. In addition to constraining each visuo-tactile embedding pair to be close to each other in latent space, it would be advantageous for surfaces with similar physical properties to be close. Other works [48, 68] have included this clustering constraint by adding an auxiliary classification objective. For many problems, semantic labels are informative of the properties that objects contain, however in our case the material labels are not always informative of tactile properties, e.g. plastics come in many forms and possess distinct surface properties but fall under one label. Yuan et al. circumvent this challenge by creating pseudo-labels formed through clustering hand-labeled tactile properties [64]. We extend unsupervised cluster labeling by creating labels from visuo-tactile embedding clusters instead of only tactile property clusters. Examples with similar tactile properties and visual statistics are encouraged to be close in space by this objective. Visuo-tactile representations are generated first by encoding features of one of the images in the sequence with a model pretrained on ImageNet [47]. The dimensionality of the feature vector is reduced through PCA and normalized to zero mean and unit variance. The reduced visual feature vector and tactile property vector are concatenated to form the final visuo-tactile representation. K-means is then used to cluster the visuo-tactile representation, creating $k$ labels.

The adversarial and classification auxiliary objectives are combined with the tactile property estimation loss and cross-modal embedding constraint to form the final cross-modal objective:

$$\mathcal{L} = \mathcal{L}_{est} + \lambda_1 \mathcal{L}_{emb} + \lambda_2 \mathcal{L}_{adv} + \lambda_3 \mathcal{L}_{class}. \tag{2}$$

**Evaluation Metric** We use the coefficient of determination ($\mathcal{R}^2$), mean absolute error (MAE), and median percentage error ($\%_{err}$) metrics to evaluate how close each estimated tactile vector is to the ground truth values. The top eight percentage error ($\%_{err}^{T8}$) is used to compare network performances on the eight best performing tactile properties in terms of $\%_{err}$. The top eight tactile properties are selected based on the metrics shown in Table 3. The $\mathcal{R}^2$ metric has been used to access the performance of vision-based regression tasks in several works [19, 53, 4, 37, 25, 5, 13]. The $\mathcal{R}^2$ metric compares the estimation from the model $\hat{t}$ against using the mean tactile value $\bar{t}$ from the training set as an estimation, and is given by:

$$\mathcal{R}^2 = 1 - \frac{\sum_i (t_i - \hat{t}_i)^2}{\sum_i (t_i - \bar{t})^2}. \tag{3}$$

### 4.2 Viewpoint Selection

**Viewpoint Selector Framework** As mentioned in Section 3, each material surface in the SPS dataset is oversampled in viewing angle space. Selectively sampling viewing angles has been shown to improve performance for action classification tasks over using all available information [71]. The challenge of selectively sampling viewing angles is formulated as follows: given a set $p$ of $N$ images collected from distinct viewing angles, select an optimal combination $q$ of $M$ images that minimize the tactile estimation error, $q^\star = \{\min \|t - f(q; w_\theta)\|_2^2 \mid q \subset p, |q| = M, |p| = N\}$. The combinatorics of this problem are too vast to explore fully, therefore we construct a sampling network $\pi(w_\pi)$ tasked with learning to select the optimal combination of viewing angles $q^\star$ based on weights $w_\pi$. The optimal combination is then used as input for a tactile estimation network $f(q^\star; w_\theta)$. We call the sampling network the Neural Viewpoint Selector (NVS) network. The NVS network is comprised of $M$ viewpoint selector vectors $z$, each tasked with choosing a single viewpoint from $N$ possible viewpoints. Each viewpoint image is assigned an equal probability of being selected. The NVS module selects a single viewing angle for the set $q$ as follows:

$$q_m = \arg\max_i \frac{exp(z_{m,i})}{\sum_{n=1}^{N} exp(z_{m,n})}, m = 1 \dots M. \tag{4}$$

The viewpoint selector vector $z$ is defined in $\mathbb{R}^N$ space. This process is repeated $M$ times with different viewpoint selector vectors to select a set of $M$ viewpoints. There are no constraints between the vectors, therefore allowing repeated viewpoints. We explore adding constraints to the selected combinations by including a value function $\mathrm{V}(q; w_v)$, which estimates how well the selected combination will perform on tactile property estimation. This network acts as a lightweight proxy for the tactile estimation network. We call this framework the Value Based Neural Viewpoint Selector (VB-NVS). The value function provides additional guidance for the viewpoint selector vectors by propagating an additional error

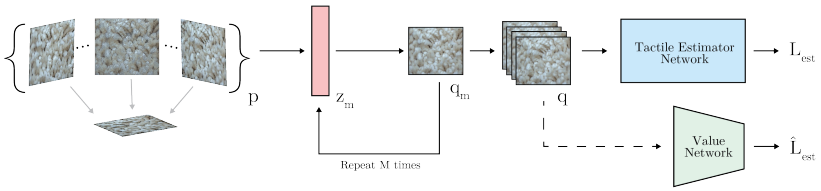


Fig. 4: **Viewpoint selection frameworks.** Inspired by recent works in neural architecture search, we construct a network (NVS) to learn which combination $q$ of viewing angles, from all available angles $p$, minimize the tactile estimator loss $L_{est}$. The NVS network is comprised of $M$ viewpoint selector vectors $z$, each responsible for selecting one viewpoint. An additional value network estimates the loss of the tactile estimator network given the selected viewpoints $q$.

signal from the objective as shown in Figure 4. The input of the value function is the combined output probability distribution from the viewpoint selector vectors. The value network then estimates the tactile estimation error $\hat{\mathcal{L}}_{est}$.

Inspired by the gradient-based neural architecture search by Pham et al. [44], the weights of the neural viewpoint selector $\pi(w_\pi)$ and tactile estimator network $f(q; w_\theta)$ are trained independently in three stages. First the weights of the $f(q; w_\theta)$ network are trained with random combinations of $q$. Then all the weights of the tactile estimator network are frozen, excluding the late-fusion layer. Combinations of viewing angles are sampled from the policy network $\pi(w_\pi)$ and evaluated with $f(q; w_\theta)$. The weights of $\pi(w_\pi)$ are updated via the gradient produced by the REINFORCE [57] objective. Finally the weights of $f(q; w_\theta)$ are reinitialized and trained with the optimal set of viewing angles $q^\star$ produced by $\pi(w_\pi)$. The VB-NVS framework is trained in a similar manner except the value network $V(q; w_v)$ is now trained in conjunction with the viewpoint selector network $\pi(w_\pi)$ and the optimal set $q^\star$ is generated by the value network instead of $\pi(w_\pi)$.

**Multi-Image Baseline** There are a variety of schemes for selecting a subset of data points. Naive approaches include random sampling and sampling at a fixed interval (equidistant). We compare our viewpoint selection networks with these naive approaches along with more advanced algorithms. Zhou et al. subsample image sequences by randomly selecting combinations of various lengths from a given image sequence [71]. The random subsamples efficiently find temporal and spatial trends from image sequences outperforming non-sampling approaches for action recognition. Rather than using a subset of viewing angles, we explore using the entire image sequence as input to the tactile property estimator. Su et al. generate feature vectors from each image separately, then fuse each feature vector together via max-pooling [50]. Early fusion methods such as I3D use 3D CNNs to learn both image-level features as well as temporal features from concatenated images [7]. For all multi-image networks except the I3D and View

Pooling, we employ late feature fusion with neural networks to aggregate features from multiple images.

**Implementation Details** The 400+ visuo-tactile pairs in the SPS dataset are randomly divided into 90/10 training/validation splits. Each experiment is rerun three times and the mean score is presented. For the single-image experiments in Section 5.1, the image corresponding to the most nadir viewing angle is selected as input. For both single-image (Section 5.1) and multi-image (Section 5.2) experiments, the mean of the five tactile measurements is used as the tactile signal. A 50-layered SE-ResNeXt [21] network pretrained on ImageNet serves as the image encoding backbone for all networks. We set the size of the latent space to be 50 and 100 for single and multi-image experiments respectively. The networks are trained for 30 epochs with a learning rate of 1e-4. Separate networks are trained for each tactile property. Non-learned and learned sampling methods select combinations of three images ($M = 3$). Additional training parameters are described in the supplementary material.

Table 2: **Single image tactile estimation ($\mathcal{R}^2$).** The $\mathcal{R}^2$ performance per tactile property is displayed, higher values are better. Our proposed cross-modal model significantly outperforms the baseline regression model across nearly all tactile properties.

| Model | fRS | cDF | tCO | cYD | aTK | mTX | cCM | cDP | cRX | mRG | mCO | uRO | tPR | uCO | fST | $\mathcal{R}^2$ | MAE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Regression | 0.07 | 0.49 | 0.50 | 0.44 | -0.46 | **0.43** | 0.13 | 0.35 | 0.11 | 0.46 | **0.56** | 0.32 | 0.57 | 0.57 | 0.53 | 0.34 | 6.17 |
| Cross-Modal | **0.54** | **0.52** | **0.62** | **0.64** | **-0.07** | **0.43** | **0.47** | **0.67** | **0.44** | **0.47** | 0.54 | **0.44** | **0.65** | **0.59** | **0.59** | **0.50** | **5.53** |

Table 3: **Single image tactile estimation ($\%_{err}$).** The median $\%_{err}$ performance per tactile property is displayed, lower values are better. Tactile properties to the left of the bold center line comprise the top eight percentage error properties.

| Model | fRS | cDF | tCO | cYD | aTK | mTX | cCM | cDP | cRX | mRG | mCO | uRO | tPR | uCO | fST | $\%_{err}$ | $\%_{err}^{T8}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Regression | 18.6 | 16.6 | 18.2 | 22.5 | **12.7** | 28.8 | 21.6 | **21.9** | 34.0 | 50.0 | 60.4 | 65.5 | 65.1 | **70.4** | 80.6 | 39.1 | 20.1 |
| Cross-Modal | **13.0** | **15.0** | **15.9** | **17.2** | 17.3 | **17.7** | **18.9** | 23.4 | **29.3** | **39.3** | **49.0** | **57.4** | **63.3** | 72.0 | **73.5** | **34.8** | **17.3** |

## 5 Experiments

### 5.1 Cross-Modal Experiments

Given a single image of a material surface, our task is to estimate the tactile properties of that surface. To highlight the effectiveness of our proposed cross-modal method, we compare the proposed method with a regression network. The results of both methods are recorded in Tables 2 and 3. Our proposed single image method outperforms the regression method across almost all fifteen tactile properties achieving better average $\mathcal{R}^2$, MAE, $\%_{err}$, and $\%_{err}^{T8}$ scores. Both networks achieve negative $\mathcal{R}^2$ scores for the adhesive tack (aTK) dimension, hence the estimates for this dimension are worse than using the average training value as a prediction. In general, the problem of estimating direct tactile properties from images-only is challenging and we expect a non-trivial margin of error.

Table 4: **Single image cross-modal ablation.** Refactoring the network as a cross-modal network with an adversarial objective greatly improves estimation performance. Visuo-tactile cluster labels outperform both material and tactile cluster labels.

| Cross-Modal | Adversarial | Material Classification | Tactile Cluster Classification [64] | Visuo-Tactile Classification | Metrics $\mathcal{R}^2$ | MAE | $\%_{err}$ | $\%^{T8}_{err}$ |
|---|---|---|---|---|---|---|---|---|
| | | | | | 0.34 | 6.17 | 39.1 | 20.1 |
| ✓ | | | | | 0.46 | 5.65 | 34.3 | 17.1 |
| ✓ | ✓ | | | | 0.49 | 5.61 | 36.2 | 18.5 |
| ✓ | | ✓ | | | 0.45 | 5.73 | 36.9 | 19.0 |
| ✓ | | | ✓ | | 0.48 | 5.60 | 35.7 | 17.7 |
| ✓ | | | | ✓ | 0.49 | 5.58 | **33.8** | **16.9** |
| ✓ | ✓ | | | ✓ | **0.50** | **5.53** | 34.8 | 17.3 |

In order to access the contribution of each component of the cross-modal network, we conduct an ablation study. In Table 4, the performance of the baseline regression network is compared to cross-modal networks with auxiliary objectives. We additionally explore using various auxiliary classification label types. As shown in Table 4, refactoring the network as a cross-modal network significantly improves the performance of the tactile estimation from an average $\mathcal{R}^2/\%_{err}$ of 0.34/39.1 to 0.46/34.3. Next, the contribution of the adversarial objective is assessed and we find that the conditional GAN objective improves the quality of generated tactile samples in terms of average $\mathcal{R}^2$ and MAE but not $\%_{err}$ or $\%^{T8}_{err}$. We then evaluate the performance of using different labels for the auxiliary latent space classification task. Using the material class labels, shown in Figure 2a, degrades the overall performance of the network. This suggests that traditional material labels do not adequately represent the tactile properties of a surface. The tactile cluster labels [64] improve results but not as much as our proposed joint visuo-tactile labels.

### 5.2 Viewing Angle Selection Experiments

After examining various network modules and frameworks for tactile property estimation from a single image, we investigate utilizing multiple images as input to the system. As described in Section 4, there are many ways of selecting a subset of images including non-learning methods such as random, equidistant, or TRN sampling and learned methods such as NVS and VB-NVS. In Tables 5 and 6, we compare various viewpoint sampling methods. Our proposed NVS and VB-NVS methods outperform all other multi-image methods in terms of average $\mathcal{R}^2$, MAE, $\%_{err}$, and $\%^{T8}_{err}$. Surprisingly, all methods that utilize the full amount of available imagery, i.e. Late Fusion, I3DNet [7], and View Pooling [50], perform much worse than the single image methods. The poor performance of the I3DNet architecture is likely a consequence of lack of significant inter-frame change in our image sequences. The View Pooling method slightly outperforms the other late fusion method. Non-learning sampling methods such as random sampling, equidistant sampling, and TRN [71] select subsamples from the total set of viewing angles without updating the selection based on performance.

The non-learned sampling methods surpass the performance of the single image model on several of the tactile properties with only equidistant sampling outperforming the single image method on average. Both NVS and VB-NVS achieve the best performance on average across all metrics while providing insightful viewpoint selection information. However, they do not outperform the single image methods in several categories suggesting that multiple images do not always provide useful information for estimating certain tactile properties. None of the multi-image methods are able to consistently provide a better than average prediction for the adhesive tack property (aTK).

Table 5: **Multi-Image tactile estimation ($\mathcal{R}^2$).** The $\mathcal{R}^2$ performance per tactile property is displayed, higher values are better. The proposed viewpoint selection frameworks outperform all other models on average. Red and blue text correspond to the first and second best scores respectively.

| Model | fRS | cDF | tCO | cYD | aTK | mTX | cCM | cDP | cRX | mRG | mCO | uRO | tPR | uCO | fST | $\mathcal{R}_2$ | MAE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Single (ours) | 0.54 | 0.52 | 0.62 | 0.64 | -0.07 | 0.43 | 0.47 | 0.67 | 0.44 | 0.47 | 0.54 | 0.44 | 0.65 | 0.59 | 0.59 | 0.50 | 5.53 |
| Late-Fusion | 0.05 | -0.06 | 0.04 | -0.21 | -0.15 | 0.46 | 0.37 | 0.04 | -0.10 | 0.31 | 0.35 | 0.04 | 0.33 | -0.01 | 0.23 | 0.11 | 9.16 |
| I3DNet [7] | 0.32 | 0.01 | -0.23 | 0.04 | -0.23 | 0.37 | 0.31 | 0.12 | -0.13 | 0.24 | 0.11 | 0.07 | 0.09 | -0.16 | 0.14 | 0.11 | 9.42 |
| View Pooling [50] | 0.03 | 0.03 | -0.01 | -0.09 | -0.11 | 0.47 | 0.23 | 0.07 | -0.05 | 0.30 | 0.32 | 0.12 | 0.25 | -0.02 | 0.10 | 0.11 | 9.01 |
| Random | 0.49 | 0.41 | 0.58 | 0.49 | -0.02 | 0.45 | 0.46 | 0.57 | 0.31 | 0.46 | 0.72 | 0.40 | 0.56 | 0.55 | 0.59 | 0.47 | 5.49 |
| Equidistant | 0.63 | 0.48 | 0.62 | 0.50 | -0.04 | 0.52 | 0.52 | 0.58 | 0.41 | 0.45 | 0.61 | 0.44 | 0.61 | 0.62 | 0.65 | 0.51 | 5.37 |
| TRN [71] | 0.64 | 0.49 | 0.52 | 0.40 | -0.10 | 0.52 | 0.55 | 0.61 | 0.35 | 0.41 | 0.66 | 0.39 | 0.53 | 0.50 | 0.63 | 0.47 | 5.53 |
| NVS (ours) | 0.62 | 0.53 | 0.63 | 0.55 | 0.02 | 0.56 | 0.53 | 0.54 | 0.49 | 0.55 | 0.68 | 0.51 | 0.54 | 0.63 | 0.64 | 0.53 | 5.34 |
| VB-NVS (ours) | 0.65 | 0.50 | 0.61 | 0.57 | -0.05 | 0.58 | 0.57 | 0.64 | 0.45 | 0.57 | 0.70 | 0.47 | 0.68 | 0.66 | 0.61 | 0.55 | 5.28 |

Table 6: **Multi-Image tactile estimation ($\%_{err}$).** The median $\%_{err}$ performance per tactile property is displayed, lower values are better. The proposed viewpoint selection frameworks outperform all other models on average. Red and blue text correspond to the first and second best scores respectively.

| Model | fRS | cDF | tCO | cYD | aTK | mTX | cCM | cDP | cRX | mRG | mCO | uRO | tPR | uCO | fST | $\%_{err}$ | $\%_{err}^{TS}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Single | 13.0 | 15.0 | 15.9 | 17.2 | 17.3 | 17.7 | 18.9 | 23.4 | 29.3 | 39.3 | 49.0 | 57.4 | 63.3 | 72.0 | 73.5 | 34.8 | 17.3 |
| Late-Fusion | 30.4 | 18.6 | 24.6 | 35.6 | 14.3 | 28.5 | 20.1 | 25.9 | 36.7 | 32.6 | 244.7 | 251.6 | 91.1 | 74.3 | 78.1 | 67.1 | 24.7 |
| I3DNet [7] | 29.9 | 19.9 | 29.4 | 25.5 | 13.9 | 24.1 | 30.6 | 36.2 | 32.7 | 54.3 | 231.1 | 131.4 | 77.7 | 85.4 | 74.2 | 59.8 | 26.2 |
| View Pooling [50] | 50.4 | 30.9 | 22.8 | 34.7 | 19.5 | 18.7 | 48.6 | 34.4 | 35.1 | 44.6 | 97.2 | 64.4 | 174.6 | 169.0 | 140.3 | 65.7 | 32.5 |
| Random | 13.2 | 18.1 | 13.7 | 17.0 | 13.4 | 15.4 | 19.2 | 22.8 | 34.7 | 38.2 | 54.7 | 62.2 | 59.3 | 59.6 | 50.8 | 32.8 | 16.6 |
| Equidistant | 16.3 | 17.2 | 12.9 | 20.7 | 9.4 | 12.9 | 19.8 | 23.3 | 35.0 | 34.8 | 52.3 | 63.0 | 60.2 | 62.4 | 58.5 | 33.2 | 16.6 |
| TRN [71] | 16.9 | 16.1 | 12.4 | 14.8 | 11.7 | 13.7 | 23.4 | 25.2 | 35.3 | 37.4 | 56.1 | 63.2 | 60.1 | 63.6 | 54.0 | 33.6 | 16.8 |
| NVS (ours) | 15.0 | 15.4 | 12.8 | 13.7 | 10.6 | 11.8 | 18.3 | 24.5 | 33.3 | 37.8 | 51.2 | 64.2 | 56.3 | 62.9 | 58.3 | 32.4 | 15.3 |
| VB-NVS (ours) | 16.3 | 13.6 | 11.5 | 12.4 | 9.6 | 17.2 | 17.6 | 26.8 | 31.8 | 32.5 | 53.2 | 59.8 | 55.1 | 67.3 | 53.9 | 31.9 | 15.6 |

In addition to improved performance from the learned subsampling methods, we gain insight into which combinations of viewing angles are useful for estimating a specific physical property. In Figure 5, the selected viewing angles from trained NVS and VB-NVS modules are shown for several tactile properties. Note, this visualization is per tactile property, not per material. The rows of Figure 5 represent the viewing angles selected for a particular tactile property while the columns represent repeated experiments. Selected viewpoints for models trained to estimate sliding resistance (fRS) are consistently close in viewing angle space for both the NVS and VB-NVS methods. The distribution of viewing angles does not vary considerably across each experiment but the location of the distribution does. This suggests that the relative difference between viewing

angles is more important for our objective than the global values of the viewing angles. The viewing angle selection is consistent with observations of prior work that angular gradients are important for material recognition [65, 62, 56]. Similar trends are observed for the macrotexture (mTX) viewing angle subsamples. The difference between the selected viewing angles for the mTX property is greater than those of the fST property suggesting that wider viewing angles are preferable to estimate macrotexture properties.

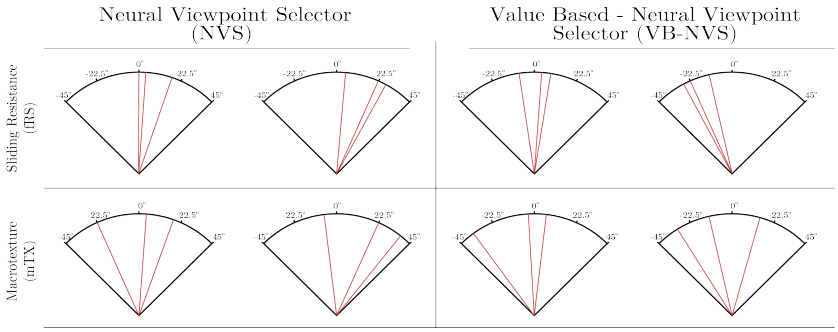


Fig. 5: **Viewpoint selection result.** The resultant selected viewpoints of both learned sampling methods. Columns represent repeated experiments, highlighting the consistency of the selected viewing angle combinations. Models optimized to estimate the sliding resistance property learn to select viewpoints that are close in viewing angle space while the selected viewpoints for the macrotexture property are farther apart.

## 6 Conclusion

This work is a pioneering step towards understanding the relationship between visual and tactile information. We propose a new challenge of estimating fifteen tactile physical properties of a surface from multiview images. We provide several methods that estimate tactile properties and determine the optimal viewing angles to sample for the estimation. To train our models we assemble the first of its kind, visuo-tactile dataset containing tactile physical properties and corresponding image sequences. We tackle the challenge of physical property estimation by designing a cross-modal network with an adversarial and a joint classification objective with results that surpass prior work in cross-modal translation. Additionally, our viewpoint selection framework achieves state-of-the-art performance for this task while providing insight as to which combinations of viewing angles are optimal for estimating a given tactile property. The proposed method can be used directly or as a prior for several tasks such as automated driving (road condition estimation), robotics (object manipulation or navigation) and manufacturing (quality control).

*Acknowledgments* This research was supported by NSF Grant #1715195. We would like to thank Eric Wengrowski, Peri Akiva, and Faith Johnson for the useful suggestions and discussions.

## References


1. Arandjelovic, R., Zisserman, A.: Objects that sound. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 435–451 (2018)
2. Arian, M.S., Blaine, C.A., Loeb, G.E., Fishel, J.A.: Using the biotac as a tumor localization tool. In: 2014 IEEE Haptics Symposium (HAPTICS). pp. 443–448. IEEE (2014)
3. Aytar, Y., Vondrick, C., Torralba, A.: See, hear, and read: Deep aligned representations. arXiv preprint arXiv:1706.00932 (2017)
4. Bessinger, Z., Jacobs, N.: Quantifying curb appeal. In: 2016 IEEE International Conference on Image Processing (ICIP). pp. 4388–4392. IEEE (2016)
5. Burgos-Artizzu, X.P., Ronchi, M.R., Perona, P.: Distance estimation of an unknown person from a portrait. In: European Conference on Computer Vision. pp. 313–327. Springer (2014)
6. Calandra, R., Owens, A., Jayaraman, D., Lin, J., Yuan, W., Malik, J., Adelson, E.H., Levine, S.: More than a feeling: Learning to grasp and regrasp using vision and touch. IEEE Robotics and Automation Letters **3**(4), 3300–3307 (2018)
7. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
8. Chen, L., Srivastava, S., Duan, Z., Xu, C.: Deep cross-modal audio-visual generation. In: Proceedings of the on Thematic Workshops of ACM Multimedia 2017. pp. 349–357. ACM (2017)
9. Chung, Y.A., Weng, W.H., Tong, S., Glass, J.: Unsupervised cross-modal alignment of speech and text embedding spaces. In: Advances in Neural Information Processing Systems. pp. 7354–7364 (2018)
10. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
11. Dai, B., Fidler, S., Urtasun, R., Lin, D.: Towards diverse and natural image descriptions via a conditional gan. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2970–2979 (2017)
12. DeGol, J., Golparvar-Fard, M., Hoiem, D.: Geometry-informed material recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1554–1562 (2016)
13. Dymczyk, M., Schneider, T., Gilitschenski, I., Siegwart, R., Stumm, E.: Erasing bad memories: Agent-side summarization for long-term mapping. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4572–4579. IEEE (2016)
14. Falco, P., Lu, S., Cirillo, A., Natale, C., Pirozzi, S., Lee, D.: Cross-modal visuotactile object recognition using robotic active exploration. In: 2017 IEEE International Conference on Robotics and Automation (ICRA). pp. 5273–5280. IEEE (2017)
15. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6202–6211 (2019)
16. Fishel, J.A., Loeb, G.E.: Bayesian exploration for intelligent identification of textures. Frontiers in neurorobotics **6**, 4 (2012)

17. Gao, Y., Hendricks, L.A., Kuchenbecker, K.J., Darrell, T.: Deep learning for tactile understanding from visual and haptic data. In: 2016 IEEE International Conference on Robotics and Automation (ICRA). pp. 536–543. IEEE (2016)
18. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research **32**(11), 1231–1237 (2013)
19. Glasner, D., Fua, P., Zickler, T., Zelnik-Manor, L.: Hot or not: Exploring correlations between appearance and temperature. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3997–4005 (2015)
20. Gonzalez-Garcia, A., van de Weijer, J., Bengio, Y.: Image-to-image translation for cross-domain disentanglement. In: Advances in Neural Information Processing Systems. pp. 1287–1298 (2018)
21. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
22. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
23. Janner, M., Levine, S., Freeman, W.T., Tenenbaum, J.B., Finn, C., Wu, J.: Reasoning about physical interactions with object-oriented prediction and planning. arXiv preprint arXiv:1812.10972 (2018)
24. Jayaraman, D., Grauman, K.: Learning to look around: Intelligently exploring unseen environments for unknown tasks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1238–1247 (2018)
25. Jean, N., Burke, M., Xie, M., Davis, W.M., Lobell, D.B., Ermon, S.: Combining satellite imagery and machine learning to predict poverty. Science **353**(6301), 790–794 (2016)
26. Jehle, M., Sommer, C., Jähne, B.: Learning of optimal illumination for material classification. In: Joint Pattern Recognition Symposium. pp. 563–572. Springer (2010)
27. Johns, E., Leutenegger, S., Davison, A.J.: Pairwise decomposition of image sequences for active multi-view recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3813–3822 (2016)
28. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3128–3137 (2015)
29. Kerr, E., McGinnity, T.M., Coleman, S.: Material recognition using tactile sensing. Expert Systems with Applications **94**, 94–111 (2018)
30. Kerzel, M., Ali, M., Ng, H.G., Wermter, S.: Haptic material classification with a multi-channel neural network. In: 2017 International Joint Conference on Neural Networks (IJCNN). pp. 439–446. IEEE (2017)
31. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539 (2014)
32. Li, R., Platt, R., Yuan, W., ten Pas, A., Roscup, N., Srinivasan, M.A., Adelson, E.: Localization and manipulation of small parts using gelsight tactile sensing. In: 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 3988–3993. IEEE (2014)
33. Li, Y., Zhu, J.Y., Tedrake, R., Torralba, A.: Connecting touch and vision via cross-modal prediction. arXiv preprint arXiv:1906.06322 (2019)
34. Liu, C., Gu, J.: Discriminative illumination: Per-pixel classification of raw materials based on optimal projections of spectral brdf. IEEE transactions on pattern analysis and machine intelligence **36**(1), 86–98 (2014)

35. Liu, C., Zoph, B., Neumann, M., Shlens, J., Hua, W., Li, L.J., Fei-Fei, L., Yuille, A., Huang, J., Murphy, K.: Progressive neural architecture search. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 19–34 (2018)
36. Liu, H., Simonyan, K., Yang, Y.: Darts: Differentiable architecture search. arXiv preprint arXiv:1806.09055 (2018)
37. McCurrie, M., Beletti, F., Parzianello, L., Westendorp, A., Anthony, S., Scheirer, W.J.: Predicting first impressions with deep learning. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). pp. 518–525. IEEE (2017)
38. Nam, H., Ha, J.W., Kim, J.: Dual attention networks for multimodal reasoning and matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 299–307 (2017)
39. Nielsen, J.B., Jensen, H.W., Ramamoorthi, R.: On optimal, minimal brdf sampling for reflectance acquisition. ACM Transactions on Graphics (TOG) **34**(6), 186 (2015)
40. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 631–648 (2018)
41. Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E.H., Freeman, W.T.: Visually indicated sounds. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2405–2413 (2016)
42. Owens, A., Wu, J., McDermott, J.H., Freeman, W.T., Torralba, A.: Ambient sound provides supervision for visual learning. In: European conference on computer vision. pp. 801–816. Springer (2016)
43. Perarnau, G., Van De Weijer, J., Raducanu, B., Álvarez, J.M.: Invertible conditional gans for image editing. arXiv preprint arXiv:1611.06355 (2016)
44. Pham, H., Guan, M.Y., Zoph, B., Le, Q.V., Dean, J.: Efficient neural architecture search via parameter sharing. arXiv preprint arXiv:1802.03268 (2018)
45. Ranjan, V., Rasiwasia, N., Jawahar, C.: Multi-label cross-modal retrieval. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4094–4102 (2015)
46. Reinecke, J., Dietrich, A., Schmidt, F., Chalon, M.: Experimental comparison of slip detection strategies by tactile sensing with the biotac® on the dlr hand arm system. In: 2014 IEEE international Conference on Robotics and Automation (ICRA). pp. 2742–2748. IEEE (2014)
47. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
48. Salvador, A., Hynes, N., Aytar, Y., Marin, J., Ofli, F., Weber, I., Torralba, A.: Learning cross-modal embeddings for cooking recipes and food images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3020–3028 (2017)
49. Shelhamer, E., Rakelly, K., Hoffman, J., Darrell, T.: Clockwork convnets for video semantic segmentation. In: European Conference on Computer Vision. pp. 852–868. Springer (2016)
50. Su, H., Maji, S., Kalogerakis, E., Learned-Miller, E.: Multi-view convolutional neural networks for 3d shape recognition. In: Proceedings of the IEEE international conference on computer vision. pp. 945–953 (2015)
51. Su, Z., Hausman, K., Chebotar, Y., Molchanov, A., Loeb, G.E., Sukhatme, G.S., Schaal, S.: Force estimation and slip detection/classification for grip control using

a biomimetic tactile sensor. In: 2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids). pp. 297–303. IEEE (2015)
52. Taylor, S., Kim, T., Yue, Y., Mahler, M., Krahe, J., Rodriguez, A.G., Hodgins, J., Matthews, I.: A deep learning approach for generalized speech animation. ACM Transactions on Graphics (TOG) **36**(4), 93 (2017)
53. Volokitin, A., Timofte, R., Van Gool, L.: Deep features or not: Temperature and time prediction in outdoor scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 63–71 (2016)
54. Wang, B., Yang, Y., Xu, X., Hanjalic, A., Shen, H.T.: Adversarial cross-modal retrieval. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 154–162. ACM (2017)
55. Wang, O., Gunawardane, P., Scher, S., Davis, J.: Material classification using brdf slices. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. pp. 2805–2811. IEEE (2009)
56. Wang, T.C., Zhu, J.Y., Hiroaki, E., Chandraker, M., Efros, A.A., Ramamoorthi, R.: A 4d light-field dataset and cnn architectures for material recognition. In: European Conference on Computer Vision. pp. 121–138. Springer (2016)
57. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine learning **8**(3-4), 229–256 (1992)
58. Wu, J., Lu, E., Kohli, P., Freeman, B., Tenenbaum, J.: Learning to see physics via visual de-animation. In: Advances in Neural Information Processing Systems. pp. 153–164 (2017)
59. Wu, J., Yildirim, I., Lim, J.J., Freeman, B., Tenenbaum, J.: Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In: Advances in neural information processing systems. pp. 127–135 (2015)
60. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3d shapenets: A deep representation for volumetric shapes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1912–1920 (2015)
61. Xu, Z., Wu, J., Zeng, A., Tenenbaum, J.B., Song, S.: Densephysnet: Learning dense physical object representations via multi-step dynamic interactions. arXiv preprint arXiv:1906.03853 (2019)
62. Xue, J., Zhang, H., Dana, K., Nishino, K.: Differential angular imaging for material recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 764–773 (2017)
63. Yuan, W., Mo, Y., Wang, S., Adelson, E.H.: Active clothing material perception using tactile sensing and deep learning. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). pp. 1–8. IEEE (2018)
64. Yuan, W., Wang, S., Dong, S., Adelson, E.: Connecting look and feel: Associating the visual and tactile properties of physical materials. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5580–5588 (2017)
65. Zhang, H., Dana, K., Nishino, K.: Reflectance hashing for material recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3071–3080 (2015)
66. Zhang, H., Dana, K., Nishino, K.: Friction from reflectance: Deep reflectance codes for predicting physical surface properties from one-shot in-field reflectance. In: Proceedings of the European Conference on Computer Vision (ECCV) (2016)
67. Zhang, H., Dana, K., Nishino, K.: Friction from reflectance: Deep reflectance codes for predicting physical surface properties from one-shot in-field reflectance. In: European Conference on Computer Vision. pp. 808–824. Springer (2016)

68. Zhang, Y., Lu, H.: Deep cross-modal projection learning for image-text matching. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 686–701 (2018)
69. Zhang, Z., Yang, L., Zheng, Y.: Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern Recognition. pp. 9242–9251 (2018)
70. Zhao, C., Sun, L., Stolkin, R.: A fully end-to-end deep learning approach for real-time simultaneous 3d reconstruction and material recognition. In: 2017 18th International Conference on Advanced Robotics (ICAR). pp. 75–82. IEEE (2017)
71. Zhou, B., Andonian, A., Oliva, A., Torralba, A.: Temporal relational reasoning in videos. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 803–818 (2018)
72. Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E.: Toward multimodal image-to-image translation. In: Advances in neural information processing systems. pp. 465–476 (2017)
73. Zoph, B., Le, Q.V.: Neural architecture search with reinforcement learning. arXiv preprint arXiv:1611.01578 (2016)