JOURNAL OF SOCIAL COMPUTING
ISSN 2688-5255 05/06 pp53–70
Volume 1, Number 1, September 2020
DOI: 10.23919/JSC.2020.0006

Darks and Stripes: Effects of Clothing on Weight Perception

Kirill Martynov, Kiran Garimella, and Robert West*

Abstract: In many societies, appearing slim (corresponding to a small body-mass index) is considered attractive.The fashion industry has been attempting to cater to this trend by designing outfits that can enhance the appearance of slimness. Two anecdotal rules, widespread in the world of fashion, are to choose dark clothes and avoid horizontal stripes, in order to appear slim. Thus far, empirical evidence has been unable to conclusively determine the validity of these rules, and there is consequently much controversy regarding the impact of both color and patterns on the visual perception of weight. In this paper, we aim to close this gap by presenting the results from a series of large-scale crowdsourcing studies that investigate the above two claims. We gathered a dataset of around 1000 images of people from the Web together with their ground-truth weight and height as well as clothing attributes about colors and patterns. To elicit the effects of colors and patterns, we asked crowd workers to estimate the weight in each image. For the analysis, we controlled potential confounds by matching images in pairs where the two images differ with respect to color or pattern, but are similar with respect to other relevant aspects. We created image pairs in two ways: firstly, observationally, i.e., from two real images; and secondly, experimentally, by manipulating the color or pattern of clothing in a real image via photo editing. Based on our analysis, we conclude that dark clothes indeed decrease perceived weight slightly but statistically significantly, and horizontal stripes have no discernible effect compared to solid light-colored clothes. These results contribute to advancing the debate around the effect of specific clothing colors and patterns and thus provide empirical grounds for everyday fashion decisions. Moreover, our work gives an outlook on the vast opportunities of using crowd sourcing in the modern fashion industry.

Key words: clothing; fashion; weight perception; body size; crowdsourcing

1 Introduction

Western female beauty standards are dominated by an ideal that favors slimness, or more technically, a small body-mass index[1,2]. Historically, the ideal body-mass index has become ever smaller over time[3]. While slimness plays a lesser role in male beauty standards, studies suggest that obesity can have a negative impact on male attractiveness[4]. Based on the widespread belief that particular choices of clothing can enhance or reduce perceived body size[5], a major question for the fashion industry as well as for individuals worldwide, revolves around the impact of clothing on body-size perception. The present paper addresses two particularly well-known pieces of anecdotal fashion advice: the claim that horizontally striped clothes increase body-size perception[6] and the claim that dark clothes decrease body-size perception[7].

On one hand, the purported advantage of dark clothes has been widely assumed in the fashion industry[8] as

• Kirill Martynov is with Google, D-80636 Munich, Germany (work done at École Polytechnique Fédérale de Lausanne (EPFL) and Technische Universität München (TUM)). E-mail: kirmartynov@gmail.com.
• Kiran Garimella is with the Massachusetts Institute of Technology (MIT), Cambridge, MA 02139, USA. E-mail: garimell@mit.edu.
• Robert West is with EPFL, CH-1004 Lausanne, Switzerland. E-mail: robert.west@epfl.ch.
∗ To whom correspondence should be addressed.
Manuscript received: 2020-08-19; accepted: 2020-09-07

well as the research community[9]. There is a general agreement that a dark object is perceived as smaller compared to a light-colored object of the same size. The assumed effect in the context of clothing is, however, mostly based on anecdotal evidence, and compared to the ubiquity of folk wisdom around dark clothing, there is a scarcity of large-scale empirical studies that could confirm or quantify the effect of dark clothes on weight perception.

On the other hand, there is much controversy about the effect of horizontal stripes. Folk wisdom states that horizontal stripes increase perceived body size[10], in stark contrast to predictions of physiological optics, such as the Helmholtz[11] and Oppel–Kundt[12, 13] illusions, which state that horizontal stripes make rectangular shapes appear both taller and thinner. Even recent empirical studies disagree about the effect of horizontal stripes: while some have claimed that horizontal stripes increased perceived weight[6], others have suggested that the difference between striped and solid-colored clothes was negligible[14], whereas Helmholtz—echoing the predictions of the aforementioned optical illusion named after him—even claimed that horizontal stripes made a figure appear taller[11]. Prior studies are, however, limited by their small scale—usually involving fewer than 100 participants, all rating a single image[6]—which makes it difficult to draw generalizable conclusions.

This work takes a novel approach to the problem, relying on crowdsourcing[15–18] to study the effect of clothing on weight perception. We conducted a series of studies in which crowd workers were shown images of people and estimated their weight and height, or indicated which of two shown people they considered to weigh less.

The image dataset consisted of around 1000 photographs depicting people wearing various styles of clothing, taken under natural circumstances and posted on an online weight-loss forum. All images were annotated with ground-truth weight and height labels by the users who uploaded them to the forum, and were annotated with color and stripe labels by the authors of this paper, using custom algorithms developed for this work.

Based on these images, we designed a matched observational study for estimating the effects of colors and stripes (Section 3.1). To overcome the limitations imposed by potential unobserved confounds, we augmented the original dataset of real images with carefully manipulated versions. In particular, starting from original images showing people wearing horizontal stripes, a graphic-design expert used Adobe Photoshop to produce versions where the originally striped clothes were replaced by solid light and solid dark clothes, respectively, while keeping everything else in the image fixed. Based on the resulting images, we designed two studies for estimating the effects of colors and stripes (Sections 3.2 and 3.3) that, due to the careful manipulation of images, are experimental in nature and thus not hampered by the same potential unobserved confounds as the observational study.

By analyzing more than 75 000 estimates from around 6500 crowd workers, we arrived at two main conclusions:

(1) Solid dark clothes indeed make a person appear to weigh slightly but statistically significantly less, such that a person switching from solid light to solid dark clothes can increase the chance of appearing to weigh less than a similar-looking reference person by 2.7 percentage points ($p = 0.0069$).

(2) Horizontal stripes and solid light colors are indistinguishable in terms of weight perception ($p = 0.58$).

In a casual but catchy formula, if D stands for solid dark, L for solid light, and S for horizontal stripes, our results about weight perception may be summarized as

$$D < L \approx S.$$

The weight of solid dark is perceived as lower than that of solid light, which is indistinguishable from that of horizontal stripes.

Taken together, this research contributes to advancing the longstanding debate around the effect of specific clothing types on body-size perception and thus helps to lay solid empirical grounds for everyday fashion advice. Given the importance of slimness for the Western ideal of beauty[2–4], the reach of our findings goes considerably beyond the world of academia. Furthermore, it showcases the vast opportunities of crowdsourcing for the modern fashion industry.

2 Annotated Image Data

We begin by describing the collection of images used in this research (Section 2.1), followed by a description of the algorithms we designed for annotating images with stripe (Section 2.2) and color (Section 2.3) labels.

2.1 Weight- and height-labeled images

The image data used in this paper were collected from a Reddit forum (“subreddit”) called *r/progresspics**,

*https://www.reddit.com/r/progresspics

where users who intend to lose (or sometimes gain) weight post pictures of themselves before and after their weight transformation. Each sample contains the ID of the Reddit post, the height and gender of the user, their weight before and after the transformation as well as one or several images. In order to be able to attach unique weight labels to all images, we automatically removed posts with more than two images ("before" and "after") and ensured that each image shows exactly one fully dressed person.

The dataset used in our analyses was assembled from two parts, A and B: Part A consists of 10 000 samples that were generously provided by Kocabey et al.[19]; Part B consists of 20 000 further samples that were crawled by the authors themselves. From Part A, we extracted 600 images (348 females, 252 males) of people wearing solid dark or solid light colors (details in Section 2.3). From the union of Parts A and B, we extracted 100 images (54 females, 46 males) of people wearing horizontally striped clothes (details in Section 2.2). Taken together, we worked with 700 annotated images.

A summary of the body-type classification by body-mass index (BMI) is presented in Table 1 (labeled "Real images"). We see that a majority of the users in our dataset are obese. Visual inspection of the dataset further revealed that around 35% of the images are full-body pictures, whereas the rest only contain upper bodies (starting from the hips), and that around 40% of the images are "selfies" taken in a mirror. The head is visible in over 90% of the images. Two samples from the dataset are shown in Fig. 1.

We automatically annotated the clothing in each image with a color attribute and a flag indicating the presence or absence of horizontal stripes. Color and stripe annotations were exclusively based on upper-body

Table 1 Classification of the images used in our studies per BMI-based body types.

(%)

Image type	Underweight	Normal	Overweight	Obese
Real images (Section 3.1)	1	18	25	56
Manipulated images (Sections 3.2 and 3.3)	1	15	24	60

Note: "Real images" refers to the images used in the observational study (Section 3.1); "manipulated images" refers to the images used in the experimental studies (Sections 3.2 and 3.3). Body-type boundaries in terms of BMI points as defined by the Centers for Disease Control and Prevention: 18.5, 25, and 30 $\mathrm{kg/m^2}$.

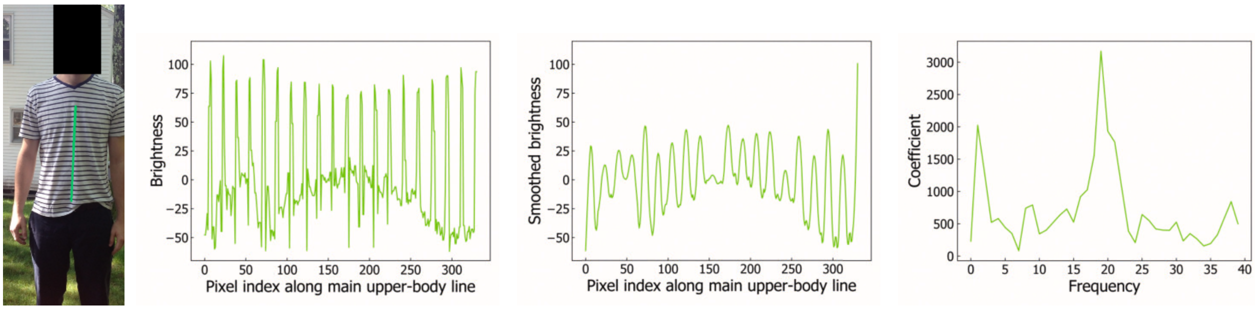
Fig. 1 Two samples from the dataset of weight- and height-labeled images collected from Reddit. Each sample contains a "before" and an "after" image. Faces censored only for paper.

clothing, for two reasons: firstly, because upper-body measurements, in particular waist circumference, are a strong indicator of obesity[20] and thus crucial for weight perception; and secondly, because only 35% of images in our dataset show full bodies (see above). The methods used to obtain stripe and color annotations are described next.

2.2 Stripe detection

To automatically detect horizontal stripes, we firstly determined the pose of the depicted body by leveraging PoseNet[21], a state-of-the-art pose detection algorithm that, given a body image, identifies all joints and body parts. Using PoseNet, we extracted the main upper-body line, running from the neck to the underbelly, and constructed a vector of brightness values along the detected line, where the brightness of a pixel is defined as the L_2-norm of the pixel's RGB triple after mean-centering by the average RGB triple along the line. Next, we remove noise by processing the brightness vector with a median filter of width 5, followed by a third-order Savitzky–Golay filter[22]. Finally, we apply a Fourier transform. The magnitude of Fourier coefficients is crucial. Our experiments suggest that horizontal stripes lead to a large coefficient for the frequency that corresponds to the actual number of visible stripes, which gives rise to the following heuristic rule: classify as horizontally striped any outfit whose maximal Fourier coefficient is associated with a frequency between 6 and 25, corresponding to a typical number of horizontal stripes, as manually determined by the authors. Figure 2 illustrates the method on an image that was classified as positive.

The above-described heuristic stripe detection algorithm has high recall but a relatively low precision of around 20%. It misclassifies as positive numerous other types of periodic variations of color, such as checkerboard patterns, letters, graphics, or shading patterns caused by certain lighting conditions. Despite



(a) Image (b) Brightness vector (c) Smoothed brightness vector (d) Fourier spectrum

Fig. 2 Steps of the stripe detection algorithm: (a) identify main upper-body line (MUBL; drawn in green); (b) construct vector of brightness values along MUBL, where brightness of a pixel equals L_2-norm of RGB triple (mean-centered along MUBL); (c) smooth the brightness vector; (d) apply Fourier transform to obtain coefficient associated with each brightness frequency. The sample is classified as positive (i.e., horizontally striped) because the frequency with the largest coefficient, 19, lies between 6 and 25. Note that exactly 19 stripes are crossed by the MUBL.

the high false-positive rate, the algorithm worked well for our purpose, as it provided a cheap and fast way to sift through more than 30 000 images and returned a set of candidates that could then be rapidly filtered by the authors by visual inspection.

2.3 Color classification

As in stripe detection, our approach to color detection starts by automatically identifying the main upper-body line using PoseNet[21]. We calculated the brightness value of an outfit as the average brightness of the pixels along the line, where the brightness of a pixel is defined as the average of the three RGB channels. That is, a brightness value of 0 corresponds to perfect black, and 255 to perfect white.

To detect solid light and solid dark clothing, we firstly picked 1000 images of people with the most clearly visible upper-body lines (i.e., images with the highest PoseNet scores for the relevant body parts), removed images with horizontal stripes (Section 2.2), split all remaining images into three approximately equally-sized groups according to their brightness value, and finally used the lower third as solid dark and the upper third as solid light. In this way, we obtained 300 images of dark-colored and 300 images of light-colored outfits. In a manual validation of 100 images, the images in the inspected sample had been perfectly classified as dark vs. light. Figure 3 shows 4 sample images with the corresponding brightness values (bv).

Fig. 3 Sample images with main upper-body line and clothing brightness value (bv). In the full dataset, these images are (from left to right) at 0%, 33%, 66%, and 100% of brightness-value.

3 Research Design

We are interested in comparing three clothing types—solid dark, solid light, and horizontally striped—with respect to their effect on perceived weight. In this section, we introduce three studies, observational as well as experimental, for estimating effects on weight perception.

3.1 Observational study: Weight estimation of real images

Our first study aimed at measuring the effect of clothing on weight perception observationally, by analyzing the crowd's weight estimates for the naturally occurring images described in Section 2.1. We firstly describe how we collected weight and height estimates via crowdsourcing and then how we matched images in pairs in order to control for potential confounds as much as possible.

Collecting weight and height estimates via crowdsourcing. We used Amazon Mechanical Turk, a popular crowdsourcing platform, to gather weight and height estimates from a diverse pool of crowd workers. We divided the set of images in the dataset (Section 2.1) into tasks with 10 images each. For each image, crowd workers guessed the weight and height of the shown person and entered their estimates into corresponding input fields located under the image. The field for weight was placed above the field for height. To provide familiar units for crowd workers from diverse backgrounds, workers could choose between kilograms and pounds for weight and between centimeters and feet/inches for height. Automatic conversion between units was performed on the fly, such that, as workers were typing their guesses in their preferred unit, the value in the other unit was updated in real time. We collected $n = 45$ independent estimates for each of the 700 images (300 light-colored, 300 dark-colored, and 100 horizontally striped outfits), for a total of more than 30 000 estimates from 3751 unique workers.

To obtain a single crowd estimate w^i_{est} for the weight of an image i, we averaged the n individual estimates via the arithmetic mean[†]:

$$w^i_{\text{est}} = \frac{1}{n}\sum_{j=1}^{n} w^i_j \tag{1}$$

where w^i_j is worker j's weight guess for image i. Similarly, we denote the average height estimate for image i as h^i_{est} and the average BMI estimate as b^i_{est}. The BMI estimate b^i_{est} was computed as $w^i_{\text{est}}/(h^i_{\text{est}})^2$ (unit: kg/m^2); it was not estimated directly by crowd workers.

The number of $n = 45$ guesses per image was determined in a pilot study where we collected the much larger number of 75 guesses for each of a small number of images and observed the convergence behavior of the mean estimates. Figure 4 illustrates for two sample images, showing that the mean stabilizes quickly, well before reaching the eventually chosen sample size of $n = 45$ guesses.

In addition to weight and height estimates or votes, we asked crowd workers to complete a survey to provide information about their own weight, height, country of residence, age, education, and gender. These statistics are summarized in Tables 2 and 3 for more than 6500 crowd workers who contributed to our three studies.

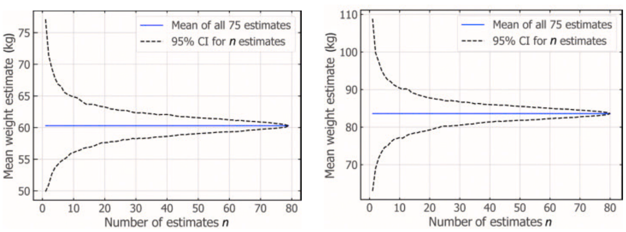


Fig. 4 Convergence of mean weight estimates for two sample images (one image left, one right) as function of number of estimates. 95% confidence intervals (CI) were computed from 1000 random permutations of all 75 estimates.

Workers were awarded \$0.01 per image. To encourage high-quality estimates, the 25% most accurate workers per image were awarded a bonus that doubled their pay for the respective image (where accuracy was computed as the absolute difference between the guess and the ground-truth weight for the image, which was available for all images as described in Section 2.1). To further increase data quality, we filtered and preprocessed the raw weight and height estimates by removing guesses from workers who did not complete the demographic survey or who appeared to have used scripts to generate random values[‡], as well as obviously erroneous guesses, such as typos or usage of wrong measurement units[§]. These data cleaning steps removed around 20% of the crowdsourced estimates.

Table 2 Weight, height, and BMI of crowd workers who contributed to the three studies.

Measure	Weight (kg)	Height (cm)	BMI (kg/m^2)
Mean	78.9	169.7	27.3
Median	74.8	169.0	25.7

Table 3 Specific distributions of crowd workers who contributed to the three studies.

Variable	Value	Relative frequency (%)
Country	United Stated	81
	India	12
	Other	7
Age	Under 21	2
	21–30	32
	31–40	34
	41–50	17
	51–60	10
	Over 60	4
Education	High-school degree	30
	Bachelor's degree	52
	Master's degree	16
	Doctor's degree	2
Gender	Male	56
	Female	44

[†]One could also use the median instead of the mean. We found the difference in results to be negligible, so we use the mean in the rest of the paper.

[‡]Concretely, we computed, for each worker, the fraction of images for which their guess ranked among the top 50%. If it happened for less than 10% of the worker's guesses, we dropped all their guesses.

[§]Concretely, we dropped guesses that were more than $z = 3$ standard deviations away from the mean guess for the respective image. The results were robust with respect to the choice of z, with $z = 2$ resulting in identical conclusions.

Matching. Our goal is to determine whether the color and pattern of clothing change the perception of weight. A naïve approach to addressing this question would be to compute the average weight estimate for each clothing type and determine whether the averages differ significantly across clothing types. The problem with this simple analysis is confounding: factors that correlate with weight estimates, such as true weight, true height, and gender[23], might also correlate with clothing choice, implying that differences between clothing types may be due to these confounding factors rather than clothing. For instance, weight estimates for women are lower than for men (because women weigh less than men on average), and if women preferred dark clothes more than men did, then if individuals wearing dark clothes appeared to weigh less, this may simply be due to the fact that women are over-represented in the dark group.

To mitigate these problems, we controlled for potential confounds by comparing people who are nearly identical with respect to a number of observed covariates while differing only with respect to clothing. Specifically, we matched images in pairs $p=(p_1,p_2)$ such that p_1 wore clothes of type c_1 (e.g., dark) and p_2 wore clothes of a different type $c_2\neq c_1$ (e.g., light), but p_1 and p_2 were as similar as possible with respect to everything else. We then define the *within-pair difference* $\Delta w_{\text{est}}^{p}$ in weight estimates as

$$\Delta w_{\text{est}}^{p}=w_{\text{est}}^{p_1}-w_{\text{est}}^{p_2} \tag{2}$$

and the average within-pair difference across all pairs $p\in P$ as

$$\Delta w_{\text{est}}=\frac{1}{|P|}\sum_{p\in P}\Delta w_{\text{est}}^{p} \tag{3}$$

Analogously, we define Δh_{est} (height) and Δb_{est} (BMI).

If all confounds were balanced by the matching—a big *if*, which led us to complement this observational study with the experimental studies introduced later—then the average within-pair difference across all pairs, Δw_{est}, would yield the size of the effect on perceived weight that clothing type c_1 affords over clothing type c_2.

Four measured covariates are available in our image dataset: gender, true weight, true height, and true BMI. We used all of them for matching, as follows: We considered a pair (p_1,p_2) of images to be a valid candidate for matching if p_1 and p_2 were of the same gender, differed in weight by at most ϵ_w, differed in height by at most ϵ_h, and differed in BMI by at most ϵ_b, but wore different types of clothes (solid light, solid dark, or horizontal stripes).

For a given pair (c_1,c_2) of clothing types, this setting can be modeled as a bipartite graph where every valid candidate pair of images is connected by an edge. To find a matching with the largest number of matched pairs, we ran an off-the-shelf maximum matching algorithm on the resulting bipartite graph.

When choosing the thresholds ϵ_w, ϵ_h, and ϵ_b, we faced a trade-off between the number of matched pairs and the quality of the resulting matching. Our choice of ϵ_b for BMI was guided by the literature, which has established 1 BMI point as the so-called *just noticeable difference*[24], i.e., the largest BMI difference that humans consistently cannot detect. We hence chose $\epsilon_b=1\ \text{kg/m}^2$. To find reasonable values for ϵ_w and ϵ_h, we ran our analysis for various values, observing that the results were robust with respect to the specific choice. We settled for $\epsilon_w=2.5\,\text{kg}$ and $\epsilon_h=2.5\,\text{cm}$. Table 4 shows that the matching process resulted in pairs of nearly identical images with respect to all observed covariates. Table 4 also contains the number of pairs created for each pairwise clothing type comparison.

3.2 Experimental study 1: Weight estimation of manipulated images

The matched observational study introduced in Section 3.1 lets us estimate the causal effect of clothing on weight perception under the condition that all confounds were balanced by the matching. As Table 4 shows, the matching did indeed balance the 4 explicitly observed confounds (true weight, true height, true BMI, and gender were balanced exactly by construction). Table 4

Table 4 Validation of pairwise image matching performed for observational study (Section 3.1), in terms of mean within-pair differences, with bootstrapped 95% confidence intervals and *p*-values from Wilcoxon's signed rank test for null hypothesis of no difference in means (i.e., large *p*-values are good, as they imply balanced pairs). Height is equal within each pair because granularity of Reddit images is 1 inch (2.54 cm), which exceeds the chosen matching threshold of $\epsilon_h=2.5\,\text{cm}$.

Classification	Number of pairs	Percentage male (%)	Actual weight difference	Actual height difference	Actual BMI difference
Dark/light	153	41	-0.04 kg $[-0.30, 0.21]$ ($p=0.71$)	0 cm ($p=1$)	-0.02 $[-0.11, 0.06]$ ($p=0.67$)
Light/striped	65	40	-0.13 kg $[-0.41, 0.36]$ ($p=0.60$)	0 cm ($p=1$)	-0.04 $[-0.14, 0.13]$ ($p=0.63$)
Dark/striped	54	39	-0.30 kg $[-0.60, 0.28]$ ($p=0.16$)	0 cm ($p=1$)	-0.10 $[-0.20, 0.09]$ ($p=0.19$)

does not, however, speak to any of the potentially large number of additional confounds that still remain and may be hard to measure because they are available only implicitly as visual information (e.g., background, face shape, body pose, camera angle, and size and fit of clothes) or that are altogether unobservable (e.g., weight awareness, fashion awareness, and mood). For instance, people in outdoor settings might both dress in certain ways and be perceived to weigh less (as outdoor settings might be associated with healthiness in raters' minds, whether consciously or not); or sharp dressers might both be more likely to wear dark (because they are more aware of the anecdotal advantages of dark clothes) and be perceived to weigh less, even when controlling for true weight.

The perfect dataset that would let us avoid such factors altogether would contain each person photographed multiple times under exactly identical conditions—including the size and fit of the clothes they wear—with only one difference: the color or pattern of the clothes they wear. In this way, all confounding factors would be eliminated, and the measured effect sizes could be attributed solely to clothing color or pattern, respectively. Unfortunately, creating such an ideal dataset would require a considerable investment, which probably explains why previous research that has adopted a similar approach worked with one single judged person[6]. Moreover, concerns regarding external validity would arise, as photographs staged for research purposes would be likely to lack the variety and naturalness of real photographs.

To circumvent these issues, we adopted a different approach: Instead of modifying the color and pattern of clothing physically at the time photographs were taken, we did so *post hoc* by manipulating photographs that had already been taken. This process is cheaper and maintains the variety and naturalness of photographs taken without experimentation in mind. Specifically, we chose, from the dataset of Section 2.1, 100 images of people wearing horizontal stripes and used Fiverr.com, an online marketplace for freelance services, to hire a graphic-design expert who manipulated each image in Adobe Photoshop by removing the horizontal stripes and producing two additional versions of the same image: one solid light, the other solid dark. The expert was paid \$5 per original image. To reduce the risk of bias, they were not informed about the purpose of the manipulated images[¶].

Two examples of real images alongside their manipulated versions are shown in Fig. 5. The example in the top row is well suited for our purposes: Even when looking closely, it is hard to tell that photo editing took place. In the example in the bottom row, on the contrary, the manipulated versions clearly look artificial. We removed such bad samples *post hoc,* after collecting the crowd estimates (as described below).

After image manipulation, experimental study 1 proceeded in exactly the same way as the observational study of Section 3.1: For each image, $n = 45$ crowd estimates were collected on Amazon Mechanical Turk[‖], and a matched analysis was performed. In contrast to

(a) (b) (c)

Fig. 5 Two samples of manipulated images as used in experimental studies (Sections 3.2 and 3.3). Starting from (a) real photographs with horizontally striped clothes, a graphic-design expert manipulated them to obtain (b) solid dark and (c) solid light versions. The manipulated versions in the top row appear realistic, whereas those in the bottom row appear artificial. Images of the latter kind were manually removed from the study.

[¶] Original and manipulated images available from the authors upon request.

[‖] Based on our experiences from the observational study, which was conducted before the experimental studies, we restricted the worker pool to residents of the United States, with the goal of avoiding country- specific biases.

the observational setup, however, the matching did not need to be performed *post hoc* in the experimental setup, since here all covariates were balanced, and thus all potential confounds controlled from the very start, by construction.

After collecting all crowd estimates, we *post hoc* addressed the issue of the low quality of some manipulated image versions by dropping all pairs of original and manipulated versions for which the estimated difference exceeded 10 kg in weight or 10 cm in height, as visual inspection revealed that in such pairs the manipulated images did not preserve the original body silhouette and were thus not well suited for our study. Although a Photoshop expert might still be able to identify the manipulated images in the remaining sample, we believe the difference is hardly noticeable for most people, especially in the context of the crowdsourcing task of height and weight estimation, on which most workers spent only a few seconds per image.

After filtering, the dataset comprised 98, 99, and 97 pairs for the dark/light, light/striped, and dark/striped comparisons, respectively. A summary of the BMI-based body type classification for the manipulated images is presented in Table 1 (labeled "Manipulated images"). The distribution of body types in the manipulated images closely matches that of the full sample of images used in the observational study.

3.3 Experimental study 2: Pairwise weight comparison of manipulated images

In both the observational study (Section 3.1) and experimental study 1 (Section 3.2), we collected absolute estimates of weight and height. Absolute weight and height estimation is, however, a hard task for humans, and it is furthermore affected strongly by personal biases on the behalf of raters[23]. Since relative judging tasks tend to be easier for humans than absolute judging tasks, we took a complementary approach in experimental study 2, asking raters to provide pairwise comparisons between images.

Recall from the previous section that, for 100 original images, we obtained manipulated versions where clothing color and pattern—and nothing else—were changed. That is, we have 100 triples where the same person is shown in light, dark, and striped clothes. Maybe the most direct way of comparing clothing types in a pairwise rating setting would be to show workers two versions of the same image and ask them in which image the person seems to weigh less. This would, however, be a highly unnatural task: Workers would realize that the weight in the two images must be identical, which would shift the focus to the meta level—"Do I think that light or dark clothing makes a person appear to weigh less?"—and lead us to measure the prevalence of anecdotal clothing advice, rather than immediate weight perception.

In our design, we therefore matched images of two *different* people into pairs, while ensuring that both images appeared similar by requiring that the estimated (not necessarily the true) weight and height were nearly identical for both people (without loss of generality, in the solid light versions of the images). Using the same maximum matching algorithm as in Section 3.1 (with $\epsilon_w = 2$ kg, $\epsilon_h = 2$ cm, and $\epsilon_b = 1$ kg/m^2), we obtained 56 matched pairs of two different people. Note that the matching ensured that every person participated in at most one pair.

In the following, we let L, D, and S stand for light, dark, and striped, respectively. Given the results of the observational study (Section 4.1) and experimental study 1 (Section 4.2), we were particularly interested in comparing light to dark clothing, $\{L, D\}$, and light to striped clothing, $\{L, S\}$. We therefore conducted experimental study 2 twice, once for $\{L, D\}$ and once for $\{L, S\}$. For ease of exposition, we shall describe the study for $\{L, D\}$, but the case $\{L, S\}$ is fully analogous.

As we have 2 images (L and D) per person, there are 4 possible *configurations* per pair of persons: *LL*, *LD*, *DL*, and *DD*. For each configuration of each person pair, $n = 40$ crowd workers guessed whether the first or the second person weighed less. (The order of the two people in a pair was randomized once and subsequently kept fixed for all rating tasks.) That is, we collected $2n = 80$ ratings for each person i in each of two clothing conditions $c \in \{L, D\}$: i wearing $c = L$ and i wearing $c = D$. Let the *vote share* $s_{ic} \in [0, 1]$ capture in what fraction of the $2n = 80$ pairwise comparisons i wearing c was judged to weigh less than the other image. The *vote-share difference*

$$\delta_i = s_{iD} - s_{iL},$$

quantifies the causal effect of clothing type on the perceived weight of person i in terms of the fraction of votes gained by wearing dark rather than light. Averaging δ_i over all people i, we obtain the overall causal effect δ of wearing dark rather than light.

In practice, each crowdsourcing task consisted of 10 comparisons, shown sequentially on separate screens. Pairs from all 4 configurations appeared in random order. Each crowd worker saw each pair in at most one of the 4 configurations, so they would not be biased by previous information. Across both runs of the experiment (for $\{L, D\}$ and $\{L, S\}$, respectively), we collected a total of 19 200 votes from 750 crowd workers.

As a quality assessment measure, most workers rated up to 4 pairs twice, which allowed us to determine test–retest reliability: 85% of workers consistently judged the same image in the pair to weigh less, a satisfactorily high value, given the hardness and subjectivity of the judging task (recall that images in a pair were explicitly chosen to be indistinguishable in terms of weight when wearing the same clothing type). As a further safeguard against random guessing, we embedded a two-letter code in each image. To enter their vote, workers had to type the corresponding code. In this way, workers had to actually look at the images—at the very least, to read the code—and could not simply click on one image randomly.

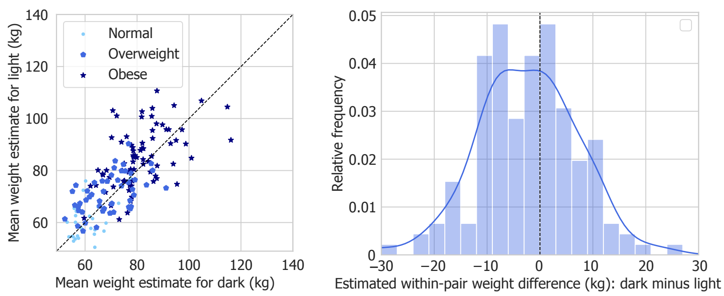
4 Result

Having described the design of the three studies, we now discuss the results for each study in turn**.

4.1 Observational study: Weight estimation of real images

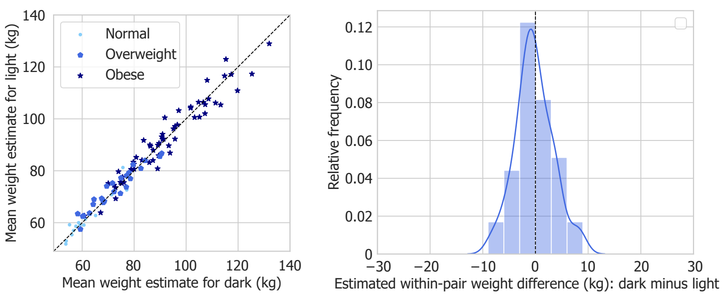
Effect of dark colors. We begin by considering the impact of clothing color on weight perception as estimated in the observational study. The weight estimates $(w^{p_1}_{\text{est}}, w^{p_2}_{\text{est}})$ (cf. Eq. (1)) for all 153 matched pairs $p = (p_1, p_2)$ are visualized in the scatter plot of Fig. 6a (left). The distribution of within-pair dark-minus-light differences $\Delta w^{p}_{\text{est}}$ in perceived weight (cf. Eq. (2)) is shown in Fig. 6a (center). Visual inspection reveals a left skew of the distribution, indicating that the dark-clad person in a pair is perceived to weigh less than the light-clad person.

Whereas the plots of Fig. 6a pertain only to the perception of weight, the first column of the table in Fig. 6a summarizes the distribution of perceived within-pair differences for all of weight, height, and BMI (Δw_{est}, Δh_{est}, and Δb_{est}; cf. Eq. (3)) in terms of averages††. We observe that average weight and BMI estimates were



	dark – light	light – striped	dark – striped
Num. pairs	153	65	54
Perceived weight diff. Δw_{est}	**−2.27 kg** [−3.74, −0.83] ($p = 0.0039$)	−0.18 kg [−2.54, 2.09] ($p = 0.68$)	**−3.08 kg** [−5.38, −0.77] ($p = 0.0072$)
Perceived height diff. Δh_{est}	−0.11 cm [−0.77, 0.55] ($p = 0.88$)	−0.21 cm [1.05, 0.64] ($p = 0.56$)	−0.01 cm [−0.97, 0.98] ($p = 0.79$)
Perceived BMI diff. Δb_{est}	**−0.77** [−1.24, −0.31] ($p = 0.0023$)	0.06 [−0.69, 0.77] ($p = 0.70$)	**−1.05** [−1.74, −0.34] ($p = 0.0071$)

(a) Observational study: Weight estimation of real images (Section 4.1)



	dark – light	light – striped	dark – striped
Num. pairs	98	99	97
Perceived weight diff. Δw_{est}	−0.23 kg [−0.93, 0.46] ($p = 0.34$)	−0.12 kg [−0.80, 0.55] ($p = 0.78$)	−0.13 kg [−0.80, 0.54] ($p = 0.50$)
Perceived height diff. Δh_{est}	0.09 cm [−0.20, 0.37] ($p = 0.58$)	0.22 cm [0.05, 0.50] ($p = 0.087$)	0.29 cm [0.01, 0.56] ($p = 0.052$)
Perceived BMI diff. Δb_{est}	−0.13 [−0.37, 0.11] ($p = 0.15$)	−0.12 [−0.35, 0.11] ($p = 0.50$)	0.17 [−0.40, 0.06] ($p = 0.061$)

(a) Experimental study 1: Weight estimation of manipulated images (Section 4.2)

Fig. 6 Results of (a) observational study (Section 4.1) and (b) experimental study (Section 4.2). Left: Scatter plot of weight estimates for all matched image pairs (dark vs. light). Center: Distribution of within-pair differences (i.e., distribution of $x - y$ from scatter plot). Right: Mean within-pair differences for weight, height, and BMI estimates, with bootstrapped 95% CIs and p-values from Wilcoxon's signed rank test (values with $p < 0.05$ in bold). For completeness, plots for all measurements (weight, height, BMI) and for comparisons of all clothing types (dark, light, striped) are available in Figs. S1–S3 (observational study) and Figs. S4–S6 (experimental study 1).

** Code: https://github.com/epfl-dlab/darks_and_stripes.

†† Moreover, plots for all of weight, height, and BMI are available in Fig. S1.

significantly lower for dark than for light, by 2.27 kg ($p = 0.004$ according to Wilcoxon's signed rank test) and 0.77 BMI points ($p = 0.002$), respectively. This implies that people who wear dark were, on average, perceived as weighing less. In contrast, height estimates were not affected by clothing color ($p = 0.88$).

Effect of horizontal stripes. Next, we investigate the effect of horizontal stripes in two matched analyses: One where we compare images with horizontal stripes to images with solid light colors (65 pairs) and one where we compare images with horizontal stripes to images with solid dark colors (54 pairs).

The results are summarized in the second and third columns of the table in Fig. 6a (also cf. Figs. S1 and S2). We observe that crowd estimates do not differ significantly between horizontally striped and solid light clothes, neither for weight nor for height nor, as a consequence, for the derived BMI ($p = 0.68$, 0.56, and 0.70, respectively). Individuals wearing dark clothes, on the contrary, were estimated as significantly less heavy ($p = 0.007$), compared to people wearing horizontal stripes, with a mean difference of 3.08 kg in favor of dark. The effect is also reflected in a difference of 1.05 BMI points ($p = 0.007$) in favor of dark.

Summary. Overall, the results from the observational study may be summarized as follows:

(1) Individuals wearing solid dark colors were judged to weigh less by a small but statistically significant amount of 2–3 kg, compared to individuals wearing solid light colors or horizontal stripes.

(2) Horizontal stripes and solid light colors were not significantly different in terms of weight perception.

(3) In terms of height perception, all three clothing types (solid light, solid dark, and horizontal stripes) were indistinguishable.

Taken together, the fact that weight, but not height, was perceived significantly lower for people wearing dark indicates that dark clothes reduce perceived body size—if the matching has balanced all confounding factors, an *if* that we remove with the experimental studies, whose results we discuss next.

4.2 Experimental study 1: Weight estimation of manipulated images

The analysis of experimental study 1 is conceptually identical to that of the observational study. The only difference between the two studies consists in the datasets used: Whereas the observational study compared images that were matched in pairs after the photographs had been taken, experimental study 1 started from images that had been created to form nearly identical pairs to begin with[**].

Effect of dark colors. As before, we visualize the weight estimates for all pairs as a scatter plot (Fig. 6b, left), display the distribution of within-pair differences of weight estimates (Fig. 6b, center), and summarize the results for all of weight, height, and BMI in a table (first column of table in Fig. 6b). We observe that the differences were much smaller in this experimental setup, compared to the observational setup (Section 4.1), presumably both because the matched images were more similar to each other and because we had retained only pairs of persons whose weight and height were judged similarly under identical clothing conditions (difference under 10 kg or 10 cm, respectively; cf. Section 3.2). Although the average within-pair difference in perceived weight between solid light and solid dark clothing had the same sign as in the observational study, the effect was much smaller (−0.23 kg vs. −2.27 kg) and not statistically significant ($p = 0.34$ according to Wilcoxon's signed rank test).

Effect of horizontal stripes. The results for the comparisons of striped with light and dark clothes are also summarized in the second and third columns of the table of Fig. 6b. As in the observational setup, light and striped were statistically indistinguishable ($p = 0.78$), and the average weight estimate for dark was smaller than for striped, but the difference was again much smaller than in the observational setup (−0.13 kg vs. −3.08 kg) and not statistically significant ($p = 0.50$).

Summary. The above results regarding weight estimation are inconclusive: On one hand, the effects point in the same direction as in the observational study, indicating that dark clothing may decrease perceived weight, but possibly due to much more closely matched image pairs, the measured effects are considerably smaller here and not statistically significant. This could either mean that there is no effect or that there is a small effect that could not be detected by the present methodology due to a small sample size. It is this disambiguity that led us to design experimental study 2, which had increased power by moving from absolute to relative weight estimation.

[**]The complete set of plots for experimental study 1 is available in Figs. S3–S5.

4.3 Experimental study 2: Pairwise weight comparison of manipulated images

In experimental study 2, we analyzed the same set of manipulated images as in experimental study 1, but using a fundamentally different methodology, based on relative pairwise weight comparison as opposed to absolute weight estimation.

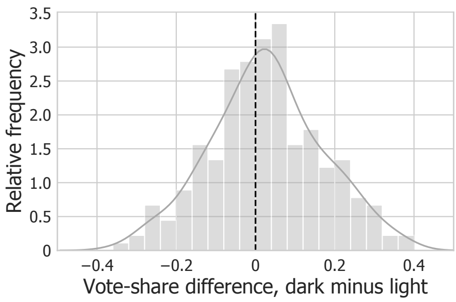
Based on the formulaic result of the observational study, $D < L \approx S$, which was qualitatively but insignificantly supported by experimental study 1, we focused on comparing solid dark to solid light clothes and solid light to horizontally striped clothes. In other words, the question was: Can we confirm that indeed $D < L$ and $L \approx S$?

Effect of dark colors. We start with the results of the comparison of dark vs. light. Recall from Section 3.3 that, for a given person i, the effect of wearing solid dark rather than solid light is captured by the vote-share difference δ_i. Averaging δ_i over all persons i, we obtain the overall causal effect δ of wearing dark, rather than light, on weight perception. It amounted to $\delta = 2.7\%$, a small but statistically significant effect ($p = 0.0069$; 95% CI [0.88%, 4.6%]). In words, when a fixed person i is compared to another, *a-priori* similar-looking person j 100 times, i can increase the number of times they are perceived to weigh less than j by 2.7 if they wear dark, rather than light.

The distribution of δ_i for the set of all people i is visualized as a histogram in Fig. 7a. The advantage of dark over light is discernible as a slight right-shift of the histogram.

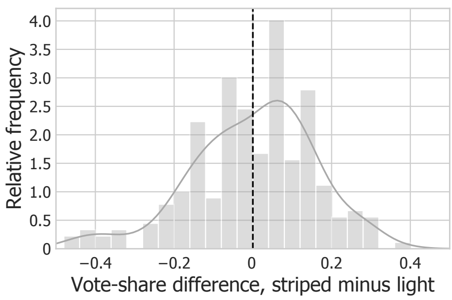
Recall that, in experimental study 2, each pair of people was rated multiple times in each of 4 configurations: *LL*, *LD*, *DL*, and *DD*. Given this structure, we may simply count, for each configuration, how often person 1 in a pair was rated as weighing less. The results, given in the right of Fig. 7a, show that both persons 1 and 2 were always perceived as weighing less when they wore dark than when they wore light: Moving vertically down, which corresponds to person 1 switching from light to dark, increases person 1's win rate; and similarly, moving horizontally right, which corresponds to person 2 switching from light to dark, increases person 2's win rate (manifested in the right of Fig. 7a as a decrease in person 1's win rate).

Note that the overall causal effect δ can also be induced from the rigiht of Fig. 7a as the average of the 4 bottom-minus-top ($D_1 L_2 - L_1 L_2$ and $D_1 D_2 - L_1 D_2$)



	L_2	D_2
L_1	0.525 [0.499, 0.540]	0.511 [0.480, 0.527]
D_1	0.566 [0.532, 0.581]	0.548 [0.520, 0.563]

(a) Dark (D) vs. light (L)



	L_2	S_2
L_1	0.513 [0.487, 0.528]	0.536 [0.495, 0.552]
S_1	0.537 [0.510, 0.551]	0.549 [0.520, 0.564]

(b) Horizontally striped (S) vs. light (L)

Fig. 7 Results of experimental study 2 (Section 4.3), comparing (a) dark (D) vs. light (L) and (b) horizontally striped (S) vs. light (L). Left: Distribution of vote-share differences ($D - L$ and $S - L$) in pairwise comparisons, where votes indicate who in the pair appears to weigh less. Right: Person 1's vote shares for all pair configurations (with 95% CIs), where rows (columns) indicate clothing type of person 1 (person 2) in pairs; values above 0.5 indicate that person 1 was estimated to weigh less than person 2 more frequently than *vice versa*.

and left-minus-right ($L_1 L_2 - L_1 D_2$ and $D_1 L_2 - D_1 D_2$) differences, which also amounts to $\delta = 2.7\%$.

Also note that the fact that all numbers in the right of Fig. 7a are slightly greater than 50% implies that there was position bias: Although the order of the two people in each pair was randomized, crowd workers on average rated the first person as weighing less. We emphasize that this does not alter our conclusions, as we work with differences of vote shares, rather than with raw vote shares directly.

Effect of horizontal stripes. Repeating the above analysis for the comparison of horizontally striped vs. solid light clothes, we cannot determine any significant effect, with an estimated overall "stripes-minus-light" effect of $\delta = 0.044\%$ ($p = 0.58$; 95% CI $[-2.0\%, 2.1\%]$).

The distribution of the individual δ_i and the number of vote shares for each of the 4 configurations are displayed in Fig. 7b.

Summary. We conclude that experimental study 2 confirmed the qualitative results from the observational study and experimental study 1, namely, that solid dark clothes make a person seem to weigh less than solid light clothes do ("$D < L$"), and that solid light clothes and horizontally striped clothes are indistinguishable in terms of weight perception ("$L \approx S$").

5 Discussion

This work is concerned with the causal effects of clothing color and patterns on perceived weight. Based on photographs taken under natural conditions and weight estimates obtained via crowdsourcing, we conducted a series of observational as well as experimental studies, all arriving at the same qualitative conclusions: Solid dark colors slightly but significantly decrease weight perception, compared to solid light colors or horizontal stripes, whereas horizontal stripes neither increase nor decrease weight perception significantly, compared to solid light colors.

We reached these conclusions using an inherently computational approach: Through a combination of Web-based image collection, crowdsourced label collection from thousands of study participants, expert image manipulation, and automated image processing, we managed to scale our studies up by an order of magnitude, compared to previous studies in this area. Only in this way did the small effect of solid dark clothes become measurable: By switching from solid light to solid dark clothes, a person can increase their chance of appearing to weigh less than a similar-looking reference person by only 2.7 percentage points on average.

On the contrary, we found no evidence of the anecdotal disadvantage of horizontal stripes: By switching from solid light to horizontally striped clothes, a person's perceived weight did not change in a statistically significant way.

At first, it may seem that horizontal stripes had a height-decreasing, rather than a weight-increasing effect: From the results of experimental study 1 (cf. table in Fig. 6b), it appears that the strongest effect present (if any) is the slightly smaller *height* estimates for horizontal stripes vs. solid colors (light and dark). Although the estimated effects are small (0.22 and 0.29 cm for light and dark, respectively) and fall slightly short of the conventional 5% significance level ($p = 0.087$ and 0.052, respectively), we investigated further by running the more powerful experimental study 2 also for height in addition to weight. The results were negative; no further support for a hypothetical shortening effect of horizontal stripes was found.

In addition to increasing the number of images and votes, we also went beyond previous studies in terms of the nature of the images. Whereas prior work had mostly focused on staged photographs that were created specifically for the respective studies[6], we started from real photographs collected from the Web, taken under natural conditions independent of the goals of our research. This data collection process made our findings more robust against idiosyncrasies that might arise with pictures taken in narrow research contexts. By further augmenting our image dataset through targeted manipulations performed by a professional graphic designer, our methodology marries the advantages of both realistic data and experimental control.

Nonetheless, it should be noted that, since the images were gleaned from an online weight loss forum, they mostly depict overweight and obese individuals (Table 1), a fact that should be taken into account when interpreting our findings. Future work should specifically investigate whether the effects are identical for individuals with an underweight or normal BMI.

Although the results of all three studies presented here were consistent, it is important to point out that the studies were not pre-registered and designed and evaluated sequentially (The data used in the observational study were originally collected for another research project[23]). In order to further decrease the likelihood of false-positive findings, we thus encourage future work to replicate our results in pre-registered studies.

In terms of further limitations, we emphasize a point made in Section 2.1, namely, this paper focused on upper-body clothing. The reasons were twofold: On one hand, a large fraction of images in our collection contain upper bodies only; on the other hand, the upper body, including the abdominal area, is anecdotally particularly important for weight perception. Future endeavors should attempt to lend data-driven support for this assumption and quantify the relative importance of upper- vs. lower-body clothing for weight perception.

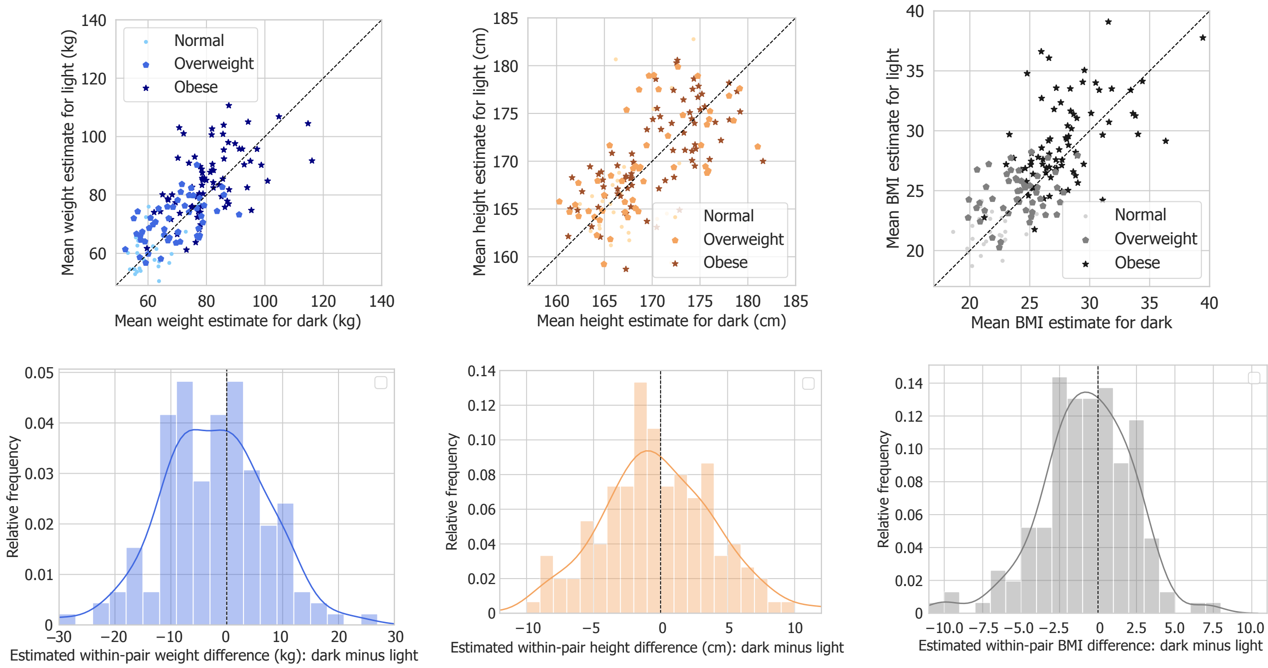
Moreover, the present research considered only horizontal stripes, foregoing the study of vertical stripes, despite the fact that folk wisdom commonly claims that vertical stripes make a person appear to weigh less. The reason for ignoring images with vertical stripes was simply that they do not occur in our dataset in sufficient numbers. Future work should make an explicit effort to collect a larger number of photographs with vertical stripes and apply our methodology to them. For instance, several large-scale fashion datasets are available[25,26], and automatically sifting them through for vertical stripes seems feasible given the computational techniques for color and stripe detection introduced in this work (Section 2).

Despite any limitations, we would like to highlight the robustness of our results, as apparent in their consistency across the three studies. Our observations are rendered considerably more reliable by the fact that each of the three studies is different in its own way: The original Reddit images are unedited and thus maximally realistic in nature, but contain numerous unmeasured covariates (e.g., background, face shape, body pose, camera angle, and size and fit of clothes), which might in the worst case explain away the effects measured in the observational study (Section 3.1). On the contrary, the manipulated images used in the experimental studies (Sections 3.2 and 3.3) are free of confounds, but were all produced by the same graphic designer and could thus conceivably be biased by consistent minor flaws, not caught by our *post-hoc* filtering. And finally, the two experimental studies were based on fundamentally different crowd labels, absolute weight guesses in the case of study 1, and relative weight comparisons in the case of study 2. The similarity of the results obtained from all three studies, each with its own strengths and weaknesses, puts our conclusions on more solid ground.

This study opens multiple interesting directions for future research. For instance, we did not attempt to answer the question of causal mechanisms. In other words, *why* do dark clothes make individuals appear to weigh slightly less? Although mathematical models have been proposed[11,14], empirical evidence sometimes contradicts the models[27]. Asking crowd workers why they think a person in an image appears to weigh less could shed new light on the question of causal pathways.

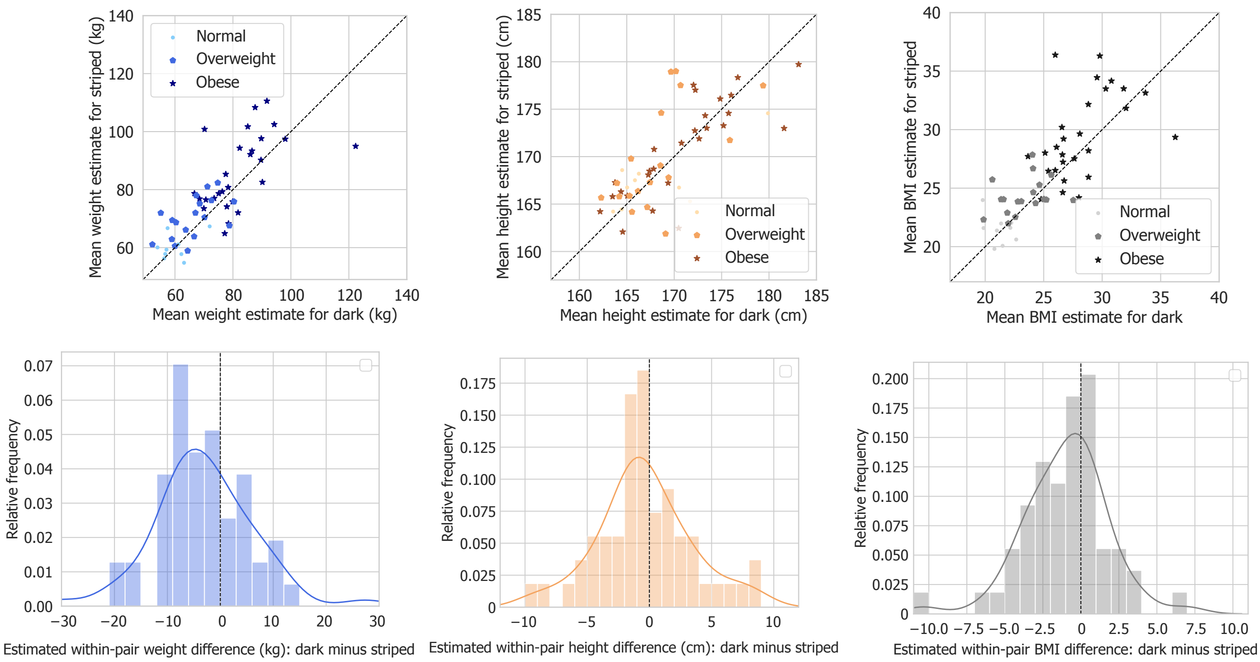
Finally, we foresee numerous practical applications of the results and techniques introduced in this work. On one hand, our results provide a solid empirical base for everyday fashion advice. On the other hand, machine-learning-driven fashion technologies[25,28] are on the rise; combining them with scalable crowdsourcing methods as leveraged in this paper could lead to tools for rapidly sensing how large populations would perceive certain clothing items.

Appendix



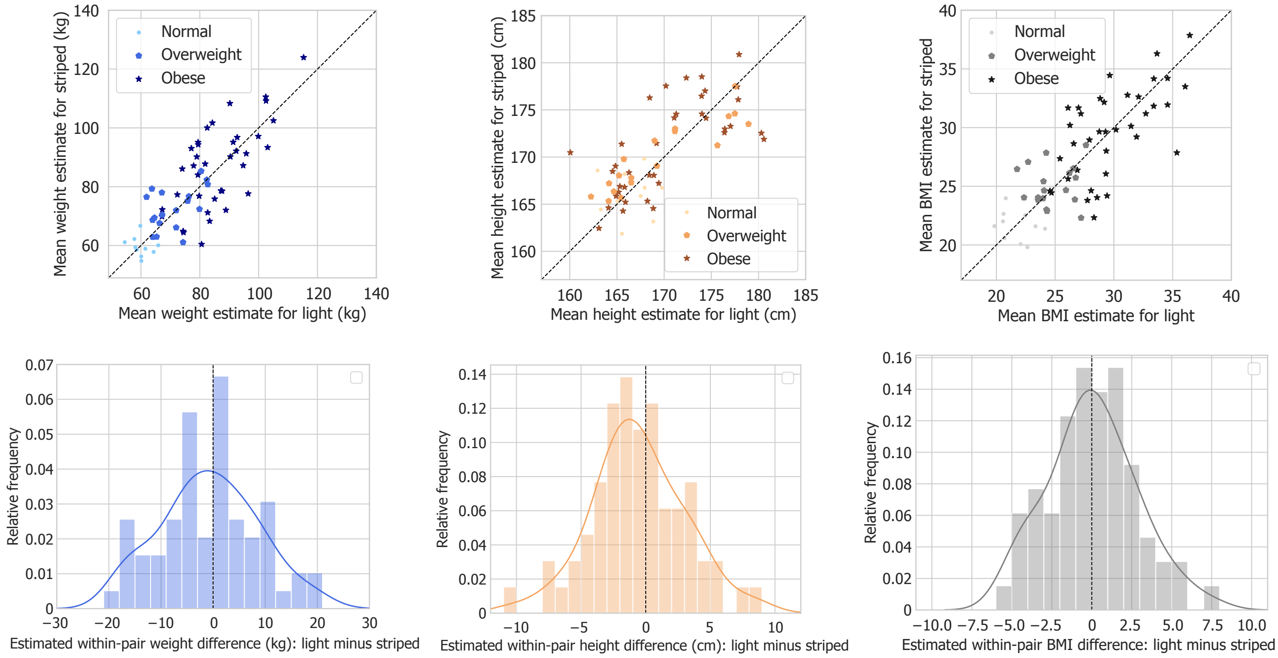
(a) Weight (b) Height (c) BMI

Fig. S1 Results of observational study, dark vs. light.



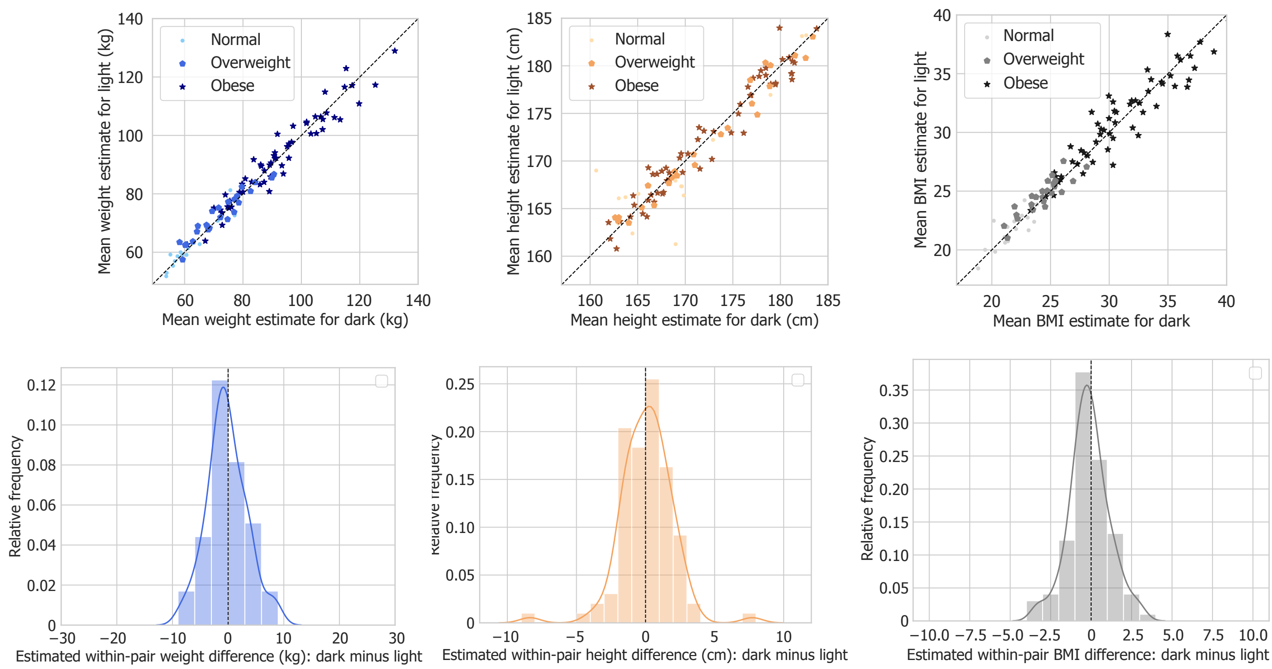
(a) Weight (b) Height (c) BMI

Fig. S2 Results of observational study, dark vs. striped.



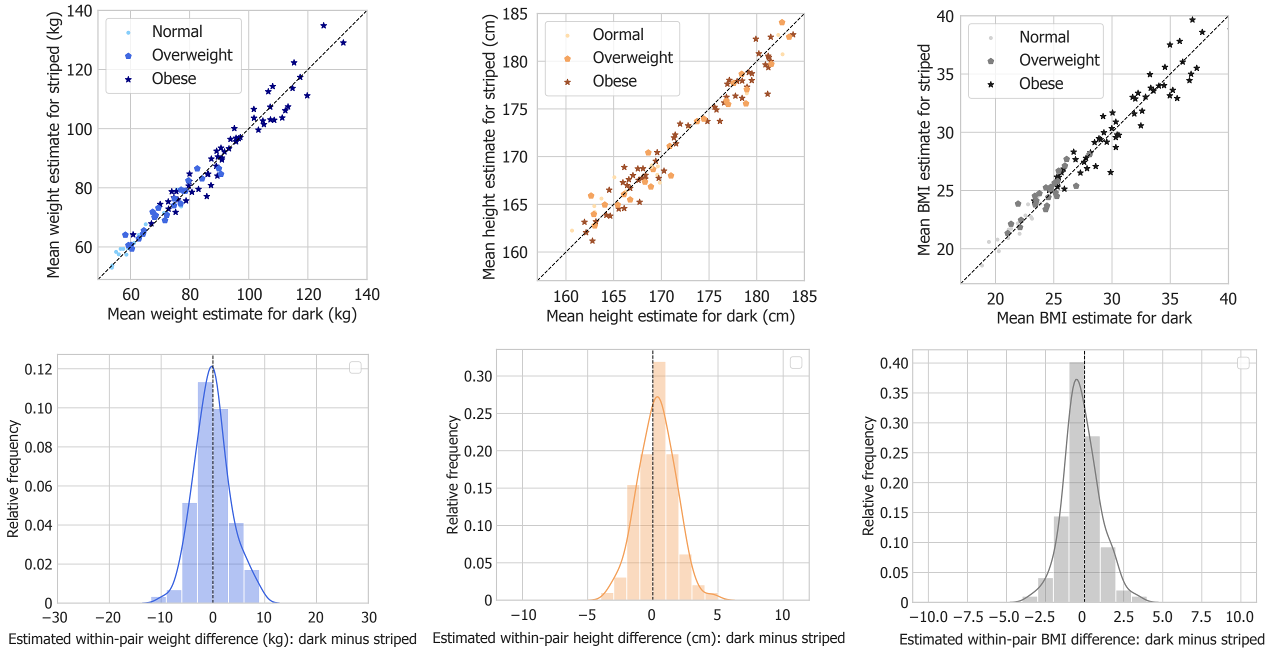
(a) Weight (b) Height (c) BMI

Fig. S3 Results of observational study, light vs. striped.



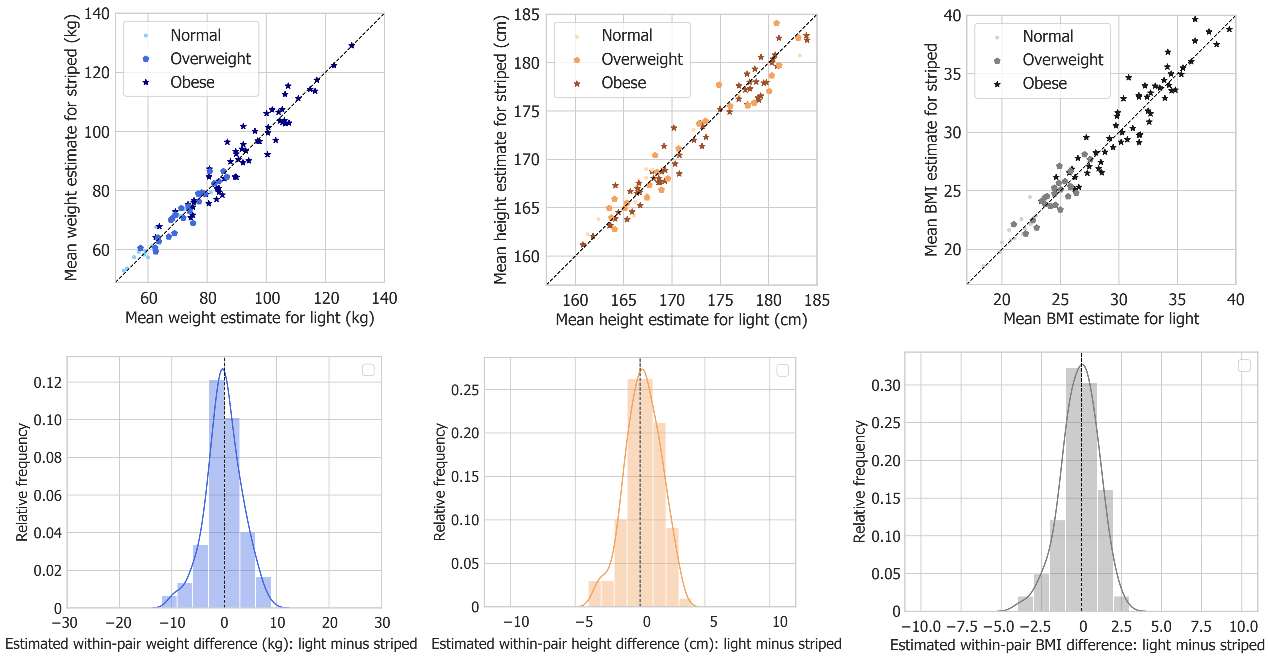
(a) Weight (b) Height (c) BMI

Fig. S4 Results of experimental study 1, dark vs. light.



(a) Weight (b) Height (c) BMI

Fig. S5 Results of experimental study 1, dark vs. striped.



(a) Weight (b) Height (c) BMI

Fig. S6 Results of experimental study 1, light vs. striped.

Acknowledgment

We thank Klaus Schönenberger, Jean-Philippe Thiran, Fernando Pérez Cruz, and their teams for valuable discussions and input. We also thank Ingmar Weber and Ferda Olfi for giving us access to their image data. This work was partly funded by a grant from the Swiss Data Science Center. West's lab is further supported by Microsoft, Google, and Facebook.

References

[1] R. Adaval, G. Saluja, and Y. W. Jiang, Seeing and thinking in pictures: A review of visual information processing, *Consumer Psychology Review*, vol. 2, no. 1, pp. 50–69, 2019.

[2] M. J. Tovée, S. Reinhardt, J. L. Emery, and P. L. Cornelissen, Optimum body-mass index and maximum sexual attractiveness, *Lancet*, vol. 352, no. 9127, p. 548, 1998.

[3] M. Voracek and M. L. Fisher, Shapely centrefolds? Temporal change in body measures: Trend analysis, *BMJ*, vol. 325, no. 7378, pp. 1447–1448, 2002.

[4] M. J. Tovée, D. S. Maisey, E. L. E. Vale, and P. L. Cornelissen, Characteristics of male attractiveness for women, *Lancet*, vol. 353, no. 9163, p. 1500, 1999.

[5] H. Frith and K. Gleeson, Clothing and embodiment: Men managing body image and appearance, *Psychology of Men & Masculinity*, vol. 5, no. 1, pp. 40–48, 2004.

[6] V. Swami and A. S. Harris, The effects of striped clothing on perceptions of body size, *Social Behavior & Personality*, vol. 40, no. 8, pp. 1239–1244, 2012.

[7] G. Winakor and R. Navarro, Effect of achromatic value of stimulus on responses to women's clothing styles, *Clothing and Textiles Research Journal*, vol. 5, no. 2, pp. 40–48, 1987.

[8] I. G. Klepp and M. Rysst, Deviant bodies and suitable clothes, *Fashion Theory*, vol. 21, no. 1, pp. 79–99, 2017.

[9] J. Kremkow, J. Z. Jin, S. J. Komban, Y. S. Wang, R. Lashgari, X. B. Li, M. Jansen, Q. Zaidi, and J. M. Alonso, Neuronal nonlinearity explains greater visual spatial resolution for darks than lights, in *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, no. 8, pp. 3170–3175, 2014.

[10] L. Feldon, *Does This Make Me Look Fat? The Definitive Rules for Dressing Thin for Every Height, Size, and Shape*. New York, NY, USA: Villard Books, 2000.

[11] H. von Helmholtz, *Treatise on Physiological Optics, Vol. III*. New York, NY, USA: Dover Publications, 1867.

[12] A. Kundt, Untersuchungen über Augenmaß und optische Täuschungen, *Annalen der Physik*, vol. 196, no. 9, pp. 118–158, 1863.

[13] N. J. Wade, D. Todorović D. Phillips, and B. Lingelbach, Johann Joseph Oppel (1855) on geometrical-optical illusions: A translation and commentary, *i-Perception*, vol. 8, no. 3, pp. 1–18, 2017.

[14] P. Thompson and K. Mikellidou, Applying the Helmholtz illusion to fashion: Horizontal stripes won't make you look fatter, *i-Perception*, vol. 2, no. 1, pp. 69–76, 2011.

[15] A. I. Chittilappilly, L. Chen, and S. Amer-Yahia, A survey of general-purpose crowdsourcing techniques, *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 9, pp. 2246–2266, 2016.

[16] A. Kittur, E. H. Chi, and B. Suh, Crowdsourcing user studies with Mechanical Turk, in *Proc. SIGCHI Conf. Human Factors in Computing Systems*, Florence, Italy, 2008, pp. 453–456.

[17] N. Luz, N. Silva, and P. Novais, A survey of task-oriented crowdsourcing, *Artificial Intelligence Review*, vol. 44, no. 2, pp. 187–213, 2015.

[18] M. C. Yuen, I. King, and K. S. Leung, A survey of crowdsourcing systems, in *Proc. IEEE 3^{rd} Int. Conf. Privacy, Security, Risk and Trust and 2011 IEEE 3^{rd} Int. Conf. Social Computing*, Boston, MA, USA, 2011, pp. 766–773.

[19] E. Kocabey, M. Camurcu, F. Ofli, Y. Aytar, J. Marin, A. Torralba, and I. Weber, Face-to-BMI: Using computer vision to infer body mass index on social media, in *Proc. 11^{th} Int. AAAI Conf. Web and Social Media*, Montreal, Canada, 2017, pp. 572–575.

[20] M. E. J. Lean, T. S. Han, and C. E. Morrison, Waist circumference as a measure for indicating need for weight management, *BMJ*, vol. 311, no. 6998, pp. 158–161, 1995.

[21] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, Realtime multi-person 2D pose estimation using part affinity fields, in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 1302-1310.

[22] A. Savitzky and M. J. E. Golay, Smoothing and differentiation of data by simplified least squares procedures, *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[23] K. Martynov, K. Garimella, and R. West, Human biases in body measurement estimation, arXiv preprint arXiv: 2009.07828, 2020.

[24] K. K. Cornelissen, L. J. Gledhill, P. L. Cornelissen, and M. J. Tovée, Visual biases in judging body weight, *British Journal of Health Psychology*, vol. 21, no. 3, pp. 555–569, 2016.

[25] Z. W. Liu, P. Luo, S. Qiu, X. G. Wang, and X. O. Tang, DeepFashion: Powering robust clothes recognition and retrieval with rich annotations, in *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 1096–1104.

[26] H. Xiao, K. Rasul, and R. Vollgraf, Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms, arXiv preprint arXiv: 1708.07747, 2017.

[27] H. Ashida, K. Kuraguchi, and K. Miyoshi, Helmholtz illusion makes you look fit only when you are already fit, but not for everyone, *i-Perception*, vol. 4, no. 5, pp. 347–351, 2013.

[28] S. Z. Zhu, S. Fidler, R. Urtasun, D. H. Lin, and C. C. Loy, Be your own Prada: Fashion synthesis with structural coherence, in *Proc. 2017 IEEE Int. Conf. Computer Vision*, Venice, Italy, 2017, pp. 1689–1697.

Kirill Martynov is a software engineer at Google. Before joining Google, he worked as a research assistant at the Data Science Lab of EPFL, Group of Mathematical Modeling of TU München, and the Corporate Technology Department of Siemens. He received the master's and bachelor's degrees in computational science from TU München.

Kiran Garimella is the Michael Hammer postdoc at MIT. Before joining MIT, he was a postdoc at EPFL. His research focuses on using digital data for social good, including areas like polarization, misinformation, and human migration. His work on studying and mitigating polarization on social media won the best student paper awards at WSDM 2017 and WebScience 2017. He received the PhD degree from Aalto University, Finland and masters & bachelors degrees from IIIT Hyderabad, India. Prior to his PhD, he worked as a research engineer at Yahoo! Research, Barcelona and QCRI, Doha.

Robert West is an assistant professor of computer science at EPFL, where he heads the Data Science Lab. His research aims to understand, predict, and enhance human behavior in social and information networks by developing techniques in computational social science, data mining, network analysis, machine learning, and natural language processing. He received the PhD degree in computer science from Stanford University, the MSc degree from McGill University, and a Dipl.-Inf. from TU München.