# Unsupervised Magnification of Posture Deviations Across Subjects

Michael Dorkenwald* Uta Büchler* Björn Ommer
HCI / IWR, Heidelberg University, Germany

## Abstract

*Analyzing human posture and precisely comparing it across different subjects is essential for accurate understanding of behavior and numerous vision applications such as medical diagnostics or sports. Motion magnification techniques help to see even small deviations in posture that are invisible to the naked eye. However, they fail when comparing subtle posture differences across individuals with diverse appearance. Keypoint-based posture estimation and classification techniques can handle large variations in appearance, but are invariant to subtle deviations in posture. We present an approach to unsupervised magnification of posture differences across individuals despite large deviations in appearance. We do not require keypoint annotation and visualize deviations on a sub-bodypart level. To transfer appearance across subjects onto a magnified posture, we propose a novel loss for disentangling appearance and posture in an autoencoder. Posture magnification yields exaggerated images that are different from the training set. Therefore, we incorporate magnification already into the training of the disentangled autoencoder and learn on real data and synthesized magnifications without supervision. Experiments confirm that our approach improves upon the state-of-the-art in magnification and on the application of discovering posture deviations due to impairment.*

## 1. Introduction

Automatic analysis of human posture, movement, and behavior is a key problem of computer vision with numerous applications such as autonomous driving [37, 57, 39, 20, 42], surveillance [11, 30, 47, 56], and health-care [7, 32, 54, 2]. A main challenge is to compare related behavior across different subjects despite vast differences in appearance. Typical approaches towards behavior understanding include action classification [8, 38, 18, 9, 3], posture estimation [1, 29, 44, 6, 5, 45] and tracking [25, 10, 1]. However, as a consequence of being invariant to appearance variability, these methods neglect subtle differences in posture. At the opposite end are motion magnification techniques [35, 55, 16, 52, 53, 50, 58, 43] that compare frames of the same video to visually amplify their subtle differences. Even recent learning-based approaches [43] are designed to magnify the intra-video differences of the same subject. Being trained for invariance to synthetic appearance changes, they can handle intra-video appearance variability, but fail at the differences across subjects and videos.

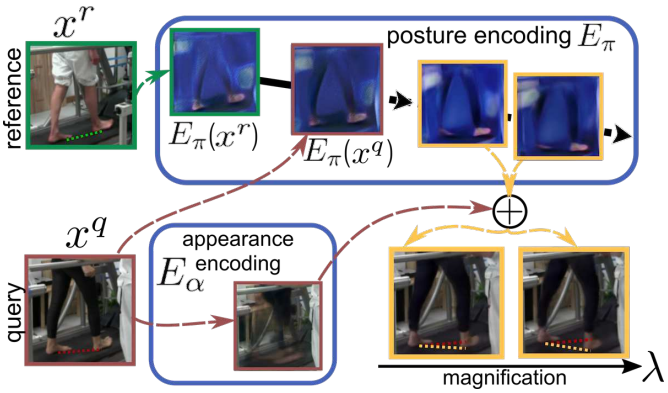


Figure 1. Magnification of Posture Deviations Across Subjects. We visually emphasize subtle posture deviations between a query $x^q$ and reference frame $x^r$ by magnifying their differences in the posture encoding. $x^q$ walks with its legs apart, highlighted by the red line in comparison to the green line of $x^r$. We first disentangle posture from appearance (the blue boxes show visualization of $E_\pi$, $E_\alpha$). Then, we extrapolate in the posture encoding the distance of $x^r$ and $x^q$ in the direction of $x^q$. The magnified images (bottom right) are generated by combining the appearance encoding of $x^q$ and the magnified posture encoding using different magnification intensities $\lambda$. The generated images allow a user to easier see differences.

What we are lacking is the best of both worlds: a microscope that can selectively amplify subtle posture differences *across subjects* while suppressing their vast deviations in appearance. Evidently, human vision is easily overwhelmed by the inter-subject appearance differences and, consequently, fails at discovering subtle posture differences across different individuals. Numerous applications ranging from sports (comparing and identifying suboptimal movement) to medicine (discovering impairment of motor behavior) would therefore benefit from such a detailed analysis.

*Indicates equal contribution

We propose an unsupervised approach to magnification of subtle posture differences across diverse subjects that requires no keypoint annotation. Being markerless, our approach can truly discover deviations and localize them on a sub-bodypart level—without having to identify the relevant body parts a priori. To transfer appearance across individuals onto the synthesized image of magnified posture differences, we explicitly disentangle posture and appearance in an autoencoder. In contrast to [43], we additionally propose a novel loss that better enforces disentanglement despite large appearance deviations. Magnification typically aims at generating new, exaggerated postures that are not in the training set and therefore difficult to synthesize. Consequently, we need to integrate the magnification process already into the training of the autoencoder. In contrast to [43] whose training works on synthetic data, our approach can directly train on inferred magnifications of real data without requiring supervision.

Experiments demonstrate that our method leads to more detailed and realistically looking magnifications. It also quantitatively improves state-of-the-art performance in terms of quality and on the downstream task of discovering posture deviation due to impairment.

The main contributions of our work are as follows: *(1)* We introduce the novel application scenario of magnifying posture deviations across subjects; *(2)* we present an unsupervised approach that separates posture from the remaining image components and enables us to directly train the magnification on real data; *(3)* we introduce three new datasets and evaluate our model on several applications.

## 2. Related Work

**Magnification.** Magnification is a valuable tool to enhance differences on images or a set of images, in order to automatically detect and visualize small deformations. Tali *et al*. [12] and Tlusty *et al*. [49] visualize non-local variations between repeating structures in a single image or for multiple views. Wadhwa *et al*. [51] exaggerates the geometric deviation between an object of interest and an ideal geometry. Video motion magnification techniques [35, 55, 16, 52, 53, 50, 58, 43] amplify the subtle motion of objects in the same video. The first attempt of motion magnification [35] computes optical flow between video frames and then amplifies every pixel separately given the optical flow information. Following works [55, 52, 16, 53, 58, 43] do not alter pixels directly, but they decompose the video into an alternative representation, e.g., by using the frequency domain. The desired motion is then selected and used to generate the image. Oh *et al*. [43] proposed the first deep learning based approach to video motion magnification using an encoder-decoder architecture. The magnification is performed by a specialized non-linear module, trained using a synthetic dataset. We also amplify differences using video frames as input, however, we amplify the deviation in posture across individuals and videos. Moreover, our approach is trained directly on the target dataset in an unsupervised manner.

**Disentanglement.** Disentangling factors of variation in an image has been proposed for more than two decades [21, 15, 19]. Recent works show successful results using deep generative neural networks [23, 24, 48, 33, 27, 17, 40, 14, 36, 28, 34, 46]. Hu *et al*. [27] proposed an unsupervised approach which separates the encoding vector into multiple chunks and forces each part to have meaningful information using an invariance objective. There is, however, no control over the characteristics of the image extracted by each chunk which is essential for magnifying posture deviations. Denton *et al*. [13] trained a pose and content encoder by exploiting temporal information of videos. The pose encoder is trained by fooling a content discriminator using an adversarial loss. Given two frames from the same video, the content encoder minimizes the distance between them. Our appearance encoder also exploits videos to be invariant to pose, however we do not need a content and pose loss. Moreover, we propose a novel disentanglement loss which enforces that both encodings contribute equally to the generation.

## 3. Approach

First, we define the problem of magnifying posture deviations. Then, we present our unsupervised approach for separating posture and appearance to assure that the magnification only alters the posture and not the appearance. Finally, we describe our method that enables us to directly train the magnification on real data.

### 3.1. Problem Definition

Given a frame $x^q$ of a query video showing a subject performing a particular action, the objective is to amplify the differences of $x^q$ to a reference frame $x^r$ and to generate the magnified image

$$x^m = m(x^q|x^r, \lambda) \tag{1}$$

with $m$ a magnification function, $\lambda$ the amplification intensity and $x^r$ a frame from a different video and subject. This problem requires a more detailed representation than a pixel space can offer. In fact, we require a model which explicitly learns an encoding space conditioned on the input image and is able to decode the encoding back to the image space. Hence, an autoencoder (AE) is the architecture of choice. An AE consists of an encoder $E$ and a decoder $D$. $E$ extracts a lower dimensional representation of the input image $x^q$, and $D$ translates the representation back to the input space by generating the reconstructed image $\hat{x}^q$. We additionally require a model that explicitly separates posture

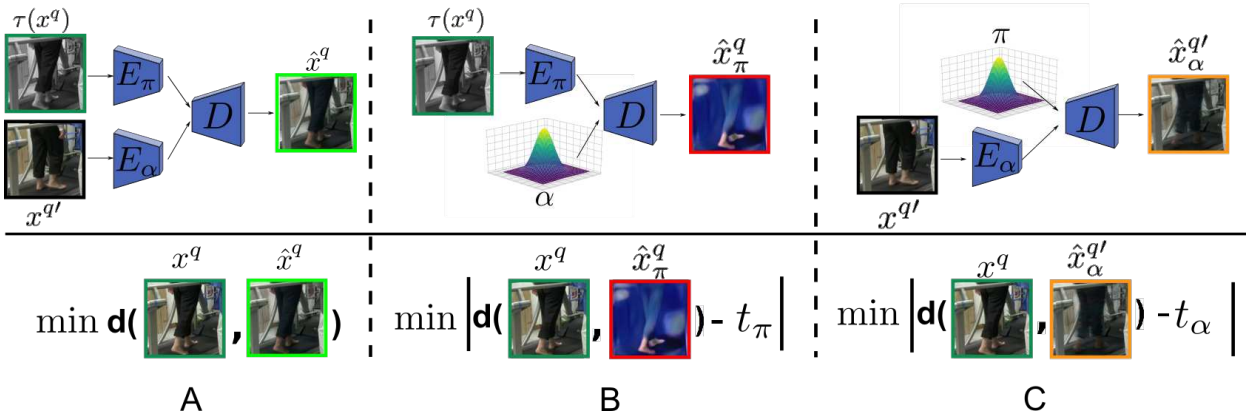


Figure 2. Disentanglement for Magnification. A: The input image $x^q$ is reconstructed by using the color transformed image $\tau(x^q)$ as input to the posture encoder and a random frame $x^{q'}$ (same video as $x^q$) as input to the appearance encoder. The reconstruction loss minimizes the distance between $x^q$ and the reconstruction $\hat{x}^q$. B and C: To enforce meaningful information in both encoders, we force the network to generate deficient images when only one of the two encodings is used. We exchange one of the two encodings with Gaussian noise, producing "fake" images and require a distance $t_\bullet$ between the original and the fake image.

from the remaining image components (appearance) since we only want to magnify the differences in posture. Hence, we use an autoencoder with two encoders, $E_\pi$ for extracting posture and $E_\alpha$ for appearance (see Fig. 2 A). The query image $x^q$ is then reconstructed as follows:

$$\hat{x}^q = D\big(E_\pi(x^q), E_\alpha(x^q)\big). \tag{2}$$

We apply the same separation for $x^r$.

Given the AE with two encoders, we can now magnify $x^q$ with respect to $x^r$ by only magnifying in the posture encoding. $x^m$ is then generated from the appearance encoding $E_\alpha(x^q)$ and the magnified posture encoding $m_\pi$ by the decoder $D$. Eq. 1 updates as follows:

$$x^m = D\Big(m_\pi\big(E_\pi(x^q)|E_\pi(x^r), \lambda\big), E_\alpha(x^q)\Big) \tag{3}$$

In the next section, we introduce our unsupervised approach to disentangle posture and appearance.

### 3.2. Disentanglement for Magnification

Magnifying posture deviations involves the comparison of subjects with different appearances. To transfer the posture from $x^r$ to $x^q$ it is crucial to obtain a posture encoding $E_\pi$ that does not contain any subject-specific information. Furthermore, we require a pure appearance representation $E_\alpha$ of $x^q$ for generating the magnified frame.

The posture and appearance encoders are considered to be disentangled if the posture encoding is invariant to appearance changes and vice-versa. The state-of-the-art in motion magnification [43] induces this invariance by introducing a regularization loss that enforces the posture representation of a color perturbed frame to be the same as the original frame. We also apply a color transformation $\tau$ to the input image $x^q$, but we additionally alter the posture by choosing a random frame $x^{q'}$ from the same video as $x^q$ ($x^{q'}$ contains the same appearance as $x^q$, but a different posture). We input $\tau(x^q)$ into the posture encoder, $x^{q'}$ into the appearance encoder and generate the reconstruction with the decoder (see Fig. 2 A). A perfect reconstruction is only possible if the AE extracts the posture information from $\tau(x^q)$ and the appearance information from $x^{q'}$. We train our model by minimizing the reconstruction loss

$$\mathcal{L}_{\text{rec}} = d(\hat{x}^q, x^q) \tag{4}$$

with $\hat{x}^q = D(E_\pi(\tau(x^q)), E_\alpha(x^{q'}))$ the reconstruction and $d(\bullet, \bullet)$ the perceptual distance [31]. Oh *et al.* [43] also employ the reconstruction loss but require an additional regularization objective to enforce invariance.

Despite the color transformation, the input $\tau(x^q)$ to the posture encoder contains appearance information such as the background scene or the type of clothes worn by the subject. This would allow the decoder to find a lazy solution by mainly leveraging $E_\pi$ to reconstruct $x^q$ as good as possible without considering the appearance encoding $E_\alpha(x^q)$. In contrast to motion magnification of single objects, the magnification of posture deviations transfers postures across subjects with different appearances. Hence, it requires a stronger separation of posture and appearance.

For that reason, we introduce a novel loss discouraging our model to correctly reproduce the image if one of the two encodings is ignored. In practice, we generate 'fake' images by exchanging the encoding of either appearance or posture with random Gaussian noise. Then we teach the network that an image reconstructed without one of the two encodings (fake image) is lacking an important component, and should therefore not be able to fully represent the original input image. We define the reconstruction with fake posture

as

$$\hat{x}_{\alpha}^{q\prime} = D\big(\mathcal{N}(0,\sigma), E_{\alpha}(x^{q\prime})\big) \tag{5}$$

and with fake appearance as

$$\hat{x}_{\pi}^{q} = D\big(E_{\pi}(\tau(x^{q})), \mathcal{N}(0,\sigma)\big). \tag{6}$$

The generation with fake images is visually illustrated in Fig. 2 B and C. We enforce a distance between the input and fake image to be close to a target value $t_{\alpha}, t_{\pi} > 0$. These values represent the lower bound on how close $\hat{x}_{\alpha}^{q\prime}$ and $\hat{x}_{\pi}^{q}$ are allowed to approach the original input $x^{q}$ during training. We update our model using the loss

$$\mathcal{L}_{\text{dis}} = \|d(x^{q}, \hat{x}_{\pi}^{q}) - t_{\pi}\|_{1} + \|d(x^{q}, \hat{x}_{\alpha}^{q\prime}) - t_{\alpha}\|_{1} \tag{7}$$

with $d(\bullet,\bullet)$ being the perceptual distance. Note, that for the distance to $\hat{x}_{\alpha}^{q\prime}$ we compare with $x^{q}$ since $x^{q}$ and $x^{q\prime}$ contain the same appearance and therefore $\hat{x}_{\alpha}^{q\prime}$ should be equal to $\hat{x}_{\alpha}^{q}$ after optimizing the network.

At this point, both terms of $\mathcal{L}_{\text{dis}}$ are optimized independently from each other and one might be easier to minimize than the other. However, to successfully generate $\hat{x}^{q}$ we require both the posture and appearance encoding to be equally advanced. Therefore, we balance the encoders by relating the target values $t_{\alpha}$ and $t_{\pi}$ with the reconstruction quality of the opposite terms,

$$t_{\pi} = d(x^{q}, \hat{x}_{\alpha}^{q\prime}) + \gamma_{\pi}, \tag{8}$$

$$t_{\alpha} = d(x^{q}, \hat{x}_{\pi}^{q}) + \gamma_{\alpha}. \tag{9}$$

with $\gamma_{\pi}$ and $\gamma_{\alpha}$ being fixed margins. If, for example, the reconstruction quality of $\hat{x}_{\alpha}^{q\prime}$ increases (and therefore the distance to $x^{q}$ decreases), $t_{\pi}$ decreases as well according to Eq. 9 and forces therefore $d(x^{q}, \hat{x}_{\pi}^{q})$ in Eq. 7 to be smaller than $\hat{x}_{\alpha}^{q\prime}$ by a margin of $\gamma_{\pi}$.

In the next section, we introduce our approach that enables us to directly train the magnification of posture deviations on the data.

### 3.3. Learning to Magnify

The magnification in the posture space usually leads to novel poses. However, it is difficult for a generative model to produce valid postures never seen during training. In particular, we require a model that is *(i)* able to precisely transfer the magnified posture $m_{\pi}$ into the image domain and *(ii)* sensitive to small encoding differences. Thus, the magnification needs to be included directly into the training process. Since ground-truth magnifications are not available, we cannot simply employ the reconstruction loss. Oh *et al.* [43] tackled this problem by creating a synthetic dataset to simulate the magnification of motion. We propose an alternative approach that allows us to directly train magnification on real data without requiring ground-truth images.

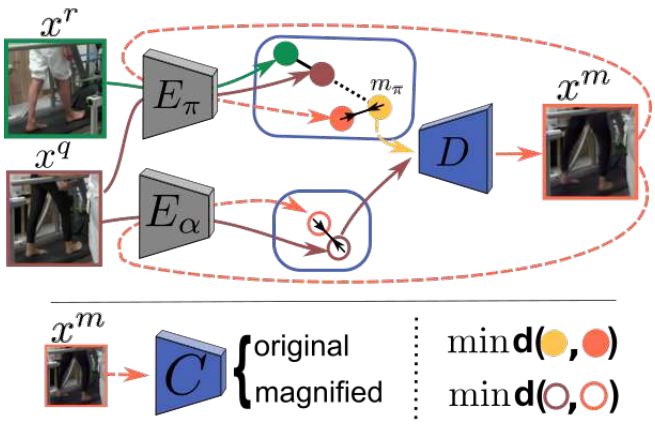


Figure 3. Learning to Magnify. Our magnification loss forces the decoder to precisely transfer the magnification $m_{\pi}$ into the image space by re-encoding the magnified image $x^{m}$ and minimizing the distance between the original magnified posture (yellow filled point) and its re-encoded posture (orange filled point). The same is applied for the appearance encodings (orange empty point and dark red empty point). Finally, an adversarial discriminator $C$ enforces the generation of realistically looking magnified images.

This way, our model produces more fine-grained and realistically looking results as demonstrated in the experimental section.

As defined in Eq. 3, we generate a magnified frame $x^{m}$ by combining the magnified posture encoding $m_{\pi}$ with the appearance encoding $E_{\alpha}(x^{q})$. For computing $m_{\pi}$ we first calculate the difference between $x^{q}$ and $x^{r}$ in the posture encoding. Then, we amplify the posture deviation in the direction of $E_{\pi}(x^{q})$. This procedure can be practically realized by extrapolating the posture differences,

$$m_{\pi}(E_{\pi}(x^{q})|E_{\pi}(x^{r}), \lambda) = E_{\pi}(x^{r}) + \lambda\,(E_{\pi}(x^{q}) - E_{\pi}(x^{r})) \tag{10}$$

with $\lambda > 1$ being the magnification factor. Fig. 1 and 3 visually depict this procedure.

During training, we require a reference frame $x^{r}$ containing a slightly different posture as $x^{q}$ since we aim to amplify subtle posture differences. This can be chosen automatically by using the $k$-th Nearest Neighbor (NN) of $x^{q}$ (excluding frames from the same video) with $k$ randomly chosen from the range $[10, 20]$. We can now generate a magnified frame $x^{m}$ for each $x^{q}$ with respect to the sampled reference frame $x^{r}$. The magnification of posture deviations requires a decoder able to precisely transfer the magnified posture encoding $m_{\pi}$ to the pixel space without distorting or losing any information about the new posture. In particular, our model should reach a fixpoint with respect to $m_{\pi}$, i.e. $m_{\pi}$ should be equal to the re-encoded decoded $m_{\pi}$,

$$m_{\pi} \overset{!}{=} E_{\pi}\big(D(m_{\pi}, \bullet)\big). \tag{11}$$

To meet this requirement, we introduce a fixpoint loss that minimizes the distance between the re-encoded magnified frame $E_\pi(x^m) = E_\pi\big(D(m_\pi, E_\alpha(x^q))\big)$ and the original magnification $m_\pi$ (see Fig. 3). We also minimize the distance between the respective appearance encodings $E_\alpha(x^m)$ and $E_\alpha(x^q)$ to ensure a consistent appearance decoding. Our model is then updated with the following fixpoint loss

$$\begin{aligned}\mathcal{L}_{\text{fix}} =& \|E_\pi(x^m) - m_\pi\big(E_\pi(x^q)|E_\pi(x^r), \lambda\big)\|_2^2 \\ &+ \|E_\alpha(x^m) - E_\alpha(x^q)\|_2^2 .\end{aligned} \tag{12}$$

We only update the decoder with $\mathcal{L}_{\text{fix}}$ since its purpose is to improve the generation of magnified images.

To encourage the decoder to produce realistically looking images, we introduce an adversarial loss. A discriminator $C$ is trained to distinguish between real images $x^q$ and magnified images $x^m$ by maximizing the adversarial loss $\mathcal{L}_{\text{A}}(C, D)$ proposed by [22],

$$\begin{aligned}\mathcal{L}_{\text{A}}(C, D) =& \mathbb{E}_{x\sim p_{\text{data}}(x)}\left[\log\, C(x)\right] \\ &+ \mathbb{E}_{\hat{x}\sim p_{\text{mag}}(x)}\left[\log\left(1 - C(\hat{x})\right]\right.\end{aligned} \tag{13}$$

with $p_{\text{data}}$ the data distribution and $p_{\text{mag}}$ the distribution of magnified images. The decoder is then trained by additionally minimizing $\mathcal{L}_A$. The adversarial loss allows us to visualize the differences in posture with higher magnification factors without generating artifacts or unrealistic images.

We summarize the losses described in this section as

$$\mathcal{L}_{\text{mag}} = \mathcal{L}_{\text{A}} + \beta\mathcal{L}_{\text{fix}} \tag{14}$$

with $\beta = 2$.

Our model is then updated with the following final loss:

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{dis}} + \gamma\mathcal{L}_{\text{mag}} \tag{15}$$

with $\gamma = 0.5$ and $\mathcal{L}_{\text{mag}}$ is only being used to update the decoder.

# 4. Experiments

We evaluate our approach on three datasets and compare our results with previous work on motion magnification. First, we introduce the datasets, perform qualitative and quantitative evaluations and demonstrate the applicability of our model on a medical scenario. Finally, we show that every component of our model is important for generating meaningful magnifications through ablation studies. The implementation details are provided in the supplementary material.

## 4.1. Datasets

Magnifying posture deviations is a challenging and new task that has never been tackled before. We propose three datasets showing three different actions for the specific task of magnifying posture deviations across subjects. It is particularly important that the datasets contain subjects with different appearances to analyze the abilities of transferring posture from one subject to another. Our datasets cover the following actions: (1) walking on a treadmill, (2) swinging a golf club and (3) moving the pupil of a person's eye.

**Human Gait Dataset to Study Disfunctional Behavior (HG2DB).** In collaboration with clinicians from University Hospital Zurich, we introduce a medical dataset for comparing postures between human subjects walking on a treadmill. The recorded patients are affected by different diseases debilitating their walking motor skills. The dataset contains also videos of healthy subjects which are used as a reference and have been recorded in the same setup. The videos display the legs of the subjects. The dataset contains 59 impaired and 10 healthy subjects with multiple recordings per subject, resulting in 229 videos with around 700 frames each, leading to a total number of 172,288 frames.

**Golf Swing.** We collected videos from Youtube showing golfers from different tournaments. The videos are recorded in slow-motion making them suitable for our scenario since many subtle differences in posture are represented. Our dataset has an overlap with the videos collected by Guha *et al*. [4] with the main difference that we use purely videos with a high frame rate. Overall, we employ 48 videos with a total number of 7000 frames. Golf Swing is more challenging than HG2DB since the videos were recorded from different tournaments (i.e., different backgrounds, lightning etc.) and they contain the full body of the person (i.e., more degrees of freedom regarding posture changes).

**Close-Up Human Eye Dataset (CUEye).** Even though eyes seem to be static if no direct movement is triggered by the person, the pupil still moves in a very subtle manner, often referred to as 'wobbling'. Magnifying posture deviations is an excellent tool to increase the visibility of this motion. We collected 10 videos showing close-up recordings of the eye from 10 different subjects (one video each) with three different eye-colors (brown, blue, green). The subjects first move their eyes to allow the generative model to differentiate between pose and appearance. This is followed by a few seconds of starring used for evaluating if our approach is able to magnify the 'wobbling' effect.

## 4.2. Qualitative Comparison

Fig. 4, 5 and 6 show magnified images generated by our model for all three datasets. We additionally provide videos in the supplementary material and on our project page[1] to further demonstrate the benefit of our magnifications.

Fig. 4 demonstrates our results on magnifying posture deviations on HG2DB (5th row; yellow border) given a reference and query frame (first row). We show the output

[1] https://compvis.github.io/magnify-posture-deviations/

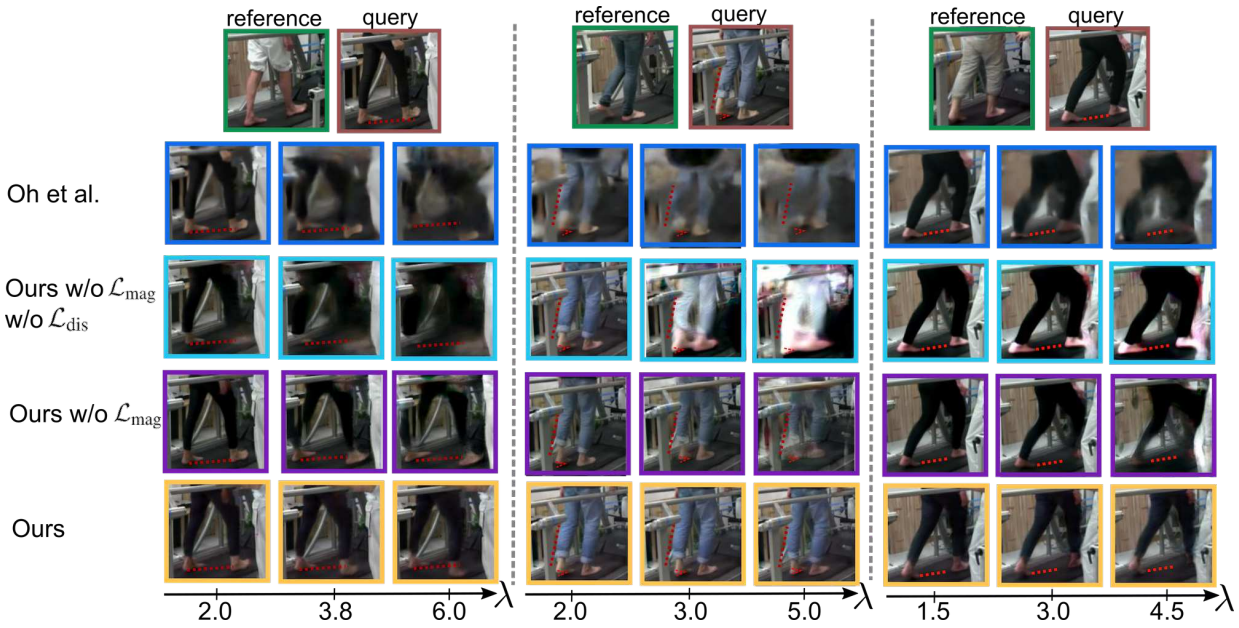


Figure 4. Qualitative Comparison on HG2DB. We show the magnification of posture deviations between a reference and query frame (first row) using the approach by Oh *et al*. [43] (2nd row), our model without $\mathcal{L}_{\text{dis}}$ and without $\mathcal{L}_{\text{mag}}$ (3rd row), our model without $\mathcal{L}_{\text{mag}}$ (4th row) and our final model (5th row). We manually superimposed red markers to facilitate the perception of the small differences and changes in the magnified images. The markers represent the posture of the query subject and are the same throughout a specific example. *Left*: The query subject keeps its legs more parallel than the reference subject. Our model exaggerates this behavior until the legs of the query subject are completely parallel. *Middle*: The query subject does not raise its left foot properly. Our magnifications visualize the differences until the left foot completely touches the treadmill. *Right*: The query subject performs bigger steps and our model further increases the distance.

with three different magnification intensities $\lambda$. Our model is able to detect the posture differences and represent the magnifications on realistically looking images.

In Fig. 5 we show our results (3rd row) on Golf Swing. Even though the dataset is very challenging due to the possible posture changes in arms and legs, our model is able to magnify the differences between the reference and query frame. In particular, the example in the middle shows that our model can even magnify arms and legs at the same time.

Fig. 6 displays the magnification of the subtle movements of a pupil while the eye is in idle state. Instead of comparing the posture deviations across different subjects, we first compute the pupil's movement in time of a query subject (top left) and transfer this motion to several target subjects with different appearances (right). Our model successfully detects the very subtle motion of the query subject and is able to transfer this motion to other subjects.

**Comparison with Previous Work.** To the best of our knowledge, this is the first approach to address the magnification of posture deviations across individuals. Previous work dealt with the task of magnifying subtle motion within the same video [35, 55, 16, 52, 53, 50, 58, 43], but not across subjects with different appearances. Considering all motion magnification approaches, [43] has the highest potential to address the more complex scenario due to their usage of a generative model with a shape and texture representation. Therefore, we qualitatively compare our results in Fig. 4 and Fig. 5 (and quantitatively in Tab. 1) with [43] on the task of magnifying posture differences. We use the official implementation of [43] from their repository. Both figures show that Oh *et al*. [43] is not specialized on magnifying posture deviations across subjects. Their model also modifies the background and appearance of the subject and therefore generates very blurry and unrealistic images. In contrast, our approach is able to precisely magnify the posture differences without altering the appearance.

### 4.3. Quantitative Analysis

**Classification of Impairment.** Our model generates novel magnified postures not present in the given dataset. Hence, we cannot directly evaluate our magnified images due to missing ground-truth magnifications. As an alternative, we introduce a quantitative evaluation based on the health condition of patients in HG2DB.

We train two linear (binary) classifiers, both on healthy vs unhealthy samples. One classifier is trained with the original images (*Original*) and the second one with the magnified generations (*Ours*). The goal is to evaluate if the magnification improves the classification of impairment and thus whether our model can support doctors during the analysis of the condition of a patient.

The classification should be subject independent and

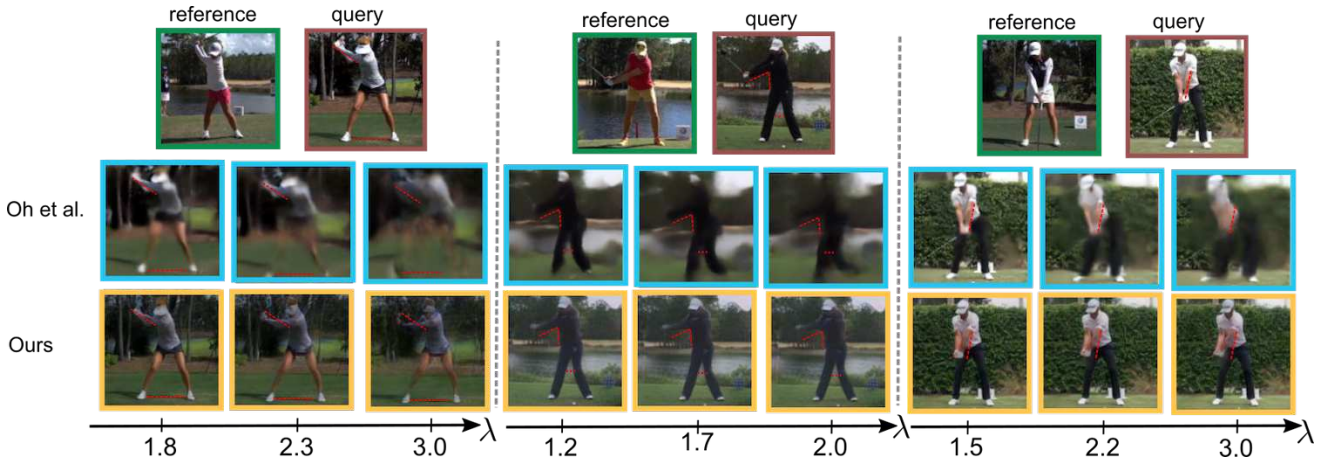


Figure 5. Qualitative Comparison on Golf Swing. We show the magnifications produced by our model and compare with previous work [43] (best viewed by zooming in on the digital version). The red markers represent the posture of the query subject and are the same throughout a specific example. *Left*: The legs of the query subject are further apart and the arm is kept lower. Our model further increases the distance of the legs and lowers the arms on the generated images. *Middle*: The right knee of the query is twisted inside and the arms are kept higher. Our approach magnifies both by further twisting the knee and raising the arms. *Right*: The reference subject is holding its arms more centered than the query subject. Our model magnifies the deviation by slowly moving the arms of the query subject to the left.

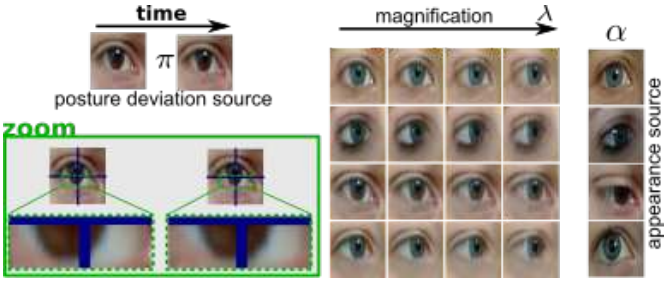


Figure 6. Magnification Results on CUEye. Detection of subtle posture differences in the pupil given a query movement (top left) and a target appearance (right). Bottom left shows a close up of the pupil with a blue grid as guide. The zoom shows a tiny motion from left to right. Given one of the target appearances shown on the right, our model can transfer the left-right movement from the query to the target appearance.

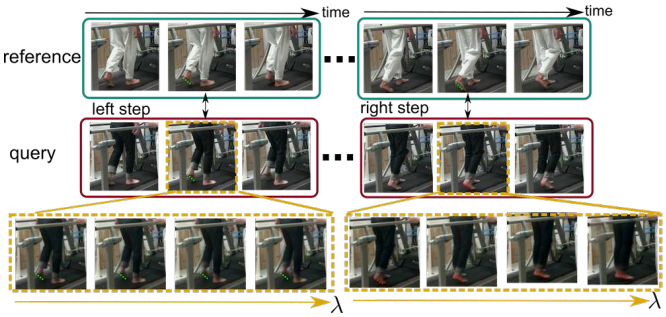


Figure 7. Magnifying Posture Deviations as Medical Tool. Deviations between healthy (first row) and impaired (second) is amplified in the generated images (third) for a better analysis of the disease status. The patient only shows difficulties during the right step. Increasing $\lambda$ emphasizes the deviation. We manually superimposed markers to facilitate the perception of the differences.

only based on posture information. For this specific experiment, we employ keypoints to represent postures since these correspond best to how humans perceive postures. In particular, we use DeepLabCut [41] for detecting the following 8 keypoints: left/right hip, left/right knee, left/right toe and left/right heel. The detector is trained with manually annotated frames of HG2DB. The keypoints are normalized using 'left hip' as origin to assure that they are comparable across different videos.

We sample 10 diverse linearly-spaced postures from a complete walking cycle sequence and perform the quantitative analysis independently per posture. We provide a visual example of the postures in the supplementary material. For every posture and subject we collect 10 Nearest Neighbors resulting in $10\times$*number of subjects* samples per posture for training and testing (in total 6900 frames). Different postures require different magnification intensities to render visible posture discrepancies between healthy and impaired subjects. Therefore, we generate the magnified images with in total 25 different magnification factors ($\lambda$), where $\lambda \in [1.2, 6]$ with a step size of $0.2$, and train one classifier per $\lambda$ and posture.

The optimal $\lambda$ per posture has been found using cross-validation and Tab. 1 reports the accuracy on the test set. We do not expect the accuracy to be $\sim 100\%$ since not all impaired patients have issues with each posture, i.e., for specific posture-subject pairs no differences to healthy subjects should be detected. This behavior can be also observed in Fig. 7. The patient (2nd row) only shows difficulties during the right step. Our model detects the deviation to an healthy

| Classifier trained with | Postures | | | | | | | | | | AVG ± STD |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |
| Original | 58.9 | 61.3 | 63.2 | 53.3 | 55.9 | 51.7 | 58.3 | 50.3 | 50.7 | 61.2 | 56.5 ± 4.5 |
| Oh *et al*. [43] | 60.2 | 61.5 | 64.1 | 53.5 | 56.1 | 52.0 | 59.4 | 52.8 | 51.2 | 61.4 | 57.2 ± 4.4 |
| Ours | **70.4** | **66.7** | **72.0** | **68.3** | **71.8** | **67.6** | **69.6** | **65.3** | **59.5** | **65.7** | **67.7 ± 3.5** |

Table 1. Classification of Impairment. We report the test accuracies (%) per posture achieved by binary classifiers (healthy vs impaired) trained and tested on the key-points of (i) the original data , (ii) the magnified images generated by previous work on motion magnification and (iii) our magnified images. A visual example of postures 1 to 10 can be found in the supplementary material.

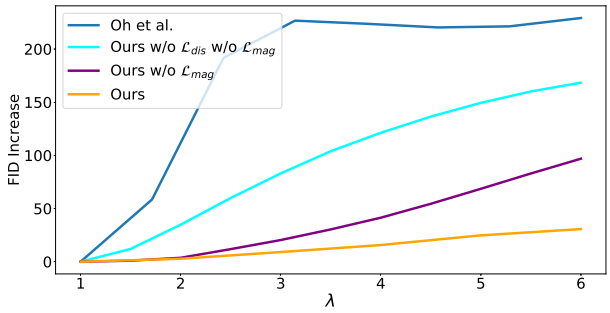


Figure 8. Quality of Visualizations - FID. We display the absolute FID increase relative to $\lambda = 1$ for different intensity values $\lambda$ using the generation of Oh *et al*. [43] (dark blue), our model without $\mathcal{L}_{\text{dis}}$ and without $\mathcal{L}_{\text{mag}}$ (light blue), our model without $\mathcal{L}_{\text{mag}}$ (purple) and our final model (orange). In contrast to our final model, the generation quality of Oh *et al*. and our incomplete models decreases significantly (FID increase) with increasing $\lambda$.

subject (1st row) and only magnifies the inaccurate posture during the right step.

We also compare our quantitative results with previous work on motion magnification. We performed the same experiment explained above using the magnified generations of Oh *et al*. [43] and report the accuracies in Tab. 1.

Most of the classifiers trained on the original data stay close to random performance and are not able to distinguish healthy from impaired. Compared to the approach of [43], our magnified images can increase the accuracy significantly. We show especially for posture 4,5 and 8 a large boost and improve the classification accuracy on average by more than 10%. This experiment demonstrates that our model is a valuable tool for discovering impairment of motor behavior.

**Quality of Visualizations - FID.** The Fréchet Inception Distance (FID) was originally proposed by Heusel *et al*. [26] and aims to evaluate the quality of generated images. It measures the distance between multivariate Gaussians for real and generated images. We have empirically found that the change in the distribution through magnification is negligible. Therefore, the FID allows us to evaluate the generation quality of our magnified images using different magnification intensities without requiring ground-truth magnifications. In Fig. 8 we show the absolute FID increase relative to $\lambda = 1$. This experiment shows that with increasing $\lambda$ our final model achieves the best results and only forfeits a small decrease in quality even with higher $\lambda$. Please note that this experiment evaluates the generation quality, not if the magnifications correspond to the actual amplification of the posture deviations.

### 4.4. Ablation Studies

In Fig. 4 and 8 we evaluate the importance of our proposed losses. We compare the magnified images produced by our full model with the generations of our model without $\mathcal{L}_{\text{mag}}$ and/or without $\mathcal{L}_{\text{dis}}$. Fig. 4 shows that our model without $\mathcal{L}_{\text{dis}}$ and $\mathcal{L}_{\text{mag}}$ generates, similar to [43], blurry and unrealistically looking images. Our model trained with the disentanglement loss $\mathcal{L}_{\text{dis}}$ improves the generations especially for smaller $\lambda$, but fails in producing valuable magnifications for larger $\lambda$. Instead, our final model is able to precisely display the magnification of posture deviations on the generated images even for large $\lambda$. Similar conclusions can be drawn from Fig. 8. The quality of the magnified images decreases for our incomplete models for $\lambda > 3$, but stays almost constant for our final model. This shows that every component of our model is important.

## 5. Conclusion

In this paper, we have introduced the problem of magnifying posture deviations across subjects and presented an approach to tackle the challenging task. Our unsupervised disentanglement allows us to only magnify posture differences while keeping the appearance unaltered. Moreover, our method enables us to integrate the magnification into the training and learn on real data without supervision. We have shown on three datasets that our approach produces valuable magnifications and improves greatly upon the performance of the state-of-the-art in motion magnification. Finally, ablation studies have demonstrated the importance of every component of our model.

## Acknowledgement

This work has been supported in part by DFG grant 421703927, the German federal ministry BMWi within the project 'KI Absicherung' and a hardware donation from NVIDIA Corporation. The authors are grateful to Linard Filli for providing the recordings used in HG2DB.

## References


[1] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5167–5176, 2018.

[2] Borislav Antic, Uta Büchler, Anna-Sophia Wahl, Martin E Schwab, and Björn Ommer. Spatiotemporal parsing of motor kinematics for assessing stroke recovery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 467–475. Springer, 2015.

[3] Borislav Antic, Timo Milbich, and Bjorn Ommer. Less is more: Video trimming for action recognition. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, June 2013.

[4] Guha Balakrishnan, Amy Zhao, Adrian V. Dalca, Frédo Durand, and John V. Guttag. Synthesizing images of humans in unseen poses. *CoRR*, abs/1804.07739, 2018.

[5] Miguel A. Bautista, Artsiom Sanakoyeu, and Bjorn Ommer. Deep unsupervised similarity learning using partially ordered sets. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[6] Miguel A Bautista, Artsiom Sanakoyeu, Ekaterina Tikhoncheva, and Bjorn Ommer. Cliquecnn: Deep unsupervised exemplar learning. In *Advances in Neural Information Processing Systems*, pages 3846–3854, 2016.

[7] Biagio Brattoli, Uta Büchler, Anna-Sophia Wahl, Martin E Schwab, and Björn Ommer. Lstm self-supervision for detailed behavior analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6466–6475, 2017.

[8] Uta Büchler, Biagio Brattoli, and Björn Ommer. Improving spatiotemporal self-supervision by deep reinforcement learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 770–786, 2018.

[9] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[10] Wongun Choi. Near-online multi-target tracking with aggregated local flow descriptor. In *Proceedings of the IEEE international conference on computer vision*, pages 3029–3037, 2015.

[11] Ankur Datta, Mubarak Shah, and N Da Vitoria Lobo. Person-on-person violence detection in video data. In *Object recognition supported by user interaction for service robots*, volume 1, pages 433–438. IEEE, 2002.

[12] Tali Dekel, Tomer Michaeli, Michal Irani, and William T. Freeman. Revealing and modifying non-local variations in a single image. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 2015.

[13] Emily Denton and Vighnesh Birodkar. Unsupervised learning of disentangled representations from video. *CoRR*, abs/1705.10915, 2017.

[14] Emily L Denton and Vighnesh Birodkar. Unsupervised learning of disentangled representations from video. In *Advances in Neural Information Processing Systems*, pages 4414–4423, 2017.

[15] Ahmed Elgammal and Chan-Su Lee. Separating style and content on a nonlinear manifold. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I. IEEE, 2004.

[16] Mohamed Elgharib, Mohamed Hefeeda, Fredo Durand, and William T Freeman. Video magnification in presence of large motions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4119–4127, 2015.

[17] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8857–8866, 2018.

[18] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6202–6211, 2019.

[19] Lloyd A. Fletcher and Rangachar Kasturi. A robust algorithm for text string separation from mixed text/graphics images. *IEEE transactions on pattern analysis and machine intelligence*, 10(6):910–918, 1988.

[20] Patrick Gebert, Alina Roitberg, Monica Haurilet, and Rainer Stiefelhagen. End-to-end prediction of driver intention using 3d convolutional neural networks. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 969–974. IEEE, 2019.

[21] Zoubin Ghahramani. Factorial learning and the em algorithm. In *Advances in neural information processing systems*, pages 617–624, 1995.

[22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[23] Naama Hadad, Lior Wolf, and Moni Shahar. A two-step disentanglement method. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 772–780, 2018.

[24] Ananya Harsh Jha, Saket Anand, Maneesh Singh, and VSR Veeravasarapu. Disentangling factors of variation with cycle-consistent variational auto-encoders. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 805–820, 2018.

[25] Roberto Henschel, Yunzhe Zou, and Bodo Rosenhahn. Multiple people tracking using body and joint detections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[26] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

[27] Qiyang Hu, Attila Szabó, Tiziano Portenier, Paolo Favaro, and Matthias Zwicker. Disentangling factors of variation

by mixing them. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3399–3407, 2018.

[28] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018.

[29] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision*, pages 34–50. Springer, 2016.

[30] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012.

[31] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[32] Mayank Kabra1, Alice A. Robie1, Marta Rivera-Alba1, Steven Branson, and Kristin Branson. Jaaba: interactive machine learning for automatic annotation of animal behavior. *Nature methods*, 10, 2012.

[33] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, et al. Fader networks: Manipulating images by sliding attributes. In *Advances in Neural Information Processing Systems*, pages 5967–5976, 2017.

[34] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018.

[35] Ce Liu, Antonio Torralba, William T Freeman, Frédo Durand, and Edward H Adelson. Motion magnification. In *ACM transactions on graphics (TOG)*, volume 24, pages 519–526. ACM, 2005.

[36] Dominik Lorenz, Leonard Bereska, Timo Milbich, and Bjorn Ommer. Unsupervised part-based disentangling of object shape and appearance. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[37] Manuel Martin, Alina Roitberg, Monica Haurilet, Matthias Horne, Simon Reiß, Michael Voit, and Rainer Stiefelhagen. Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2801–2810, 2019.

[38] Brais Martinez, Davide Modolo, Yuanjun Xiong, and Joseph Tighe. Action recognition with spatial-temporal discriminative filter banks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5482–5491, 2019.

[39] Sarfaraz Masood, Abhinav Rai, Aakash Aggarwal, Mohammad Najmud Doja, and Musheer Ahmad. Detecting distraction of drivers using convolutional neural network. *Pattern Recognition Letters*, 2018.

[40] Michael F Mathieu, Junbo Jake Zhao, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. Disentangling factors of variation in deep representation using adversarial training. In *Advances in Neural Information Processing Systems*, pages 5040–5048, 2016.

[41] Alexander Mathis, Pranav Mamidanna, Kevin M. Cury, Taiga Abe, Venkatesh N. Murthy, Mackenzie W. Mathis, and Matthias Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 2018.

[42] Sadegh Mohammadi, Alessandro Perina, Hamed Kiani, and Vittorio Murino. Angry crowds: Detecting violent events in videos. In *European Conference on Computer Vision*, pages 3–18. Springer, 2016.

[43] Tae-Hyun Oh, Ronnachai Jaroensri, Changil Kim, Mohamed Elgharib, Fr'edo Durand, William T Freeman, and Wojciech Matusik. Learning-based video motion magnification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 633–648, 2018.

[44] Umer Rafi, Bastian Leibe, Juergen Gall, and Ilya Kostrikov. An efficient convolutional network for human pose estimation. In *BMVC*, volume 1, page 2, 2016.

[45] Artsiom Sanakoyeu, Miguel A Bautista, and Björn Ommer. Deep unsupervised learning of visual similarities. *Pattern Recognition*, 78:331–343, 2018.

[46] Artsiom Sanakoyeu, Dmytro Kotovenko, Sabine Lang, and Bjorn Ommer. A style-aware content loss for real-time hd style transfer. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[47] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6479–6488, 2018.

[48] Attila Szabó, Qiyang Hu, Tiziano Portenier, Matthias Zwicker, and Paolo Favaro. Challenges in disentangling independent factors of variation. *arXiv preprint arXiv:1711.02245*, 2017.

[49] Tal Tlusty, Tomer Michaeli, Tali Dekel, and Lihi Zelnik-Manor. Modifying non-local variations across multiple views. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[50] Sergey Tulyakov, Xavier Alameda-Pineda, Elisa Ricci, Lijun Yin, Jeffrey F Cohn, and Nicu Sebe. Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2396–2404, 2016.

[51] Neal Wadhwa, Tali Dekel, Donglai Wei, Frédo Durand, and William T. Freeman. Deviation magnification: Revealing departure from ideal geometries. *ACM Trans. Graph. (Proceedings SIGGRAPH Asia 2015)*, 34(6), 2015.

[52] Neal Wadhwa, Michael Rubinstein, Frédo Durand, and William T Freeman. Phase-based video motion processing. *ACM Transactions on Graphics (TOG)*, 32(4):80, 2013.

[53] Neal Wadhwa, Michael Rubinstein, Frédo Durand, and William T Freeman. Riesz pyramids for fast phase-based video magnification. In *2014 IEEE International Conference*

*on Computational Photography (ICCP)*, pages 1–10. IEEE, 2014.

[54] Anna-Sophia Wahl, U Büchler, A Brändli, Biagio Brattoli, Simon Musall, Hansjörg Kasper, Benjamin V Ineichen, Fritjof Helmchen, Björn Ommer, and Martin E Schwab. Optogenetically stimulating intact rat corticospinal tract post-stroke restores motor control through regionalized functional circuit formation. *Nature communications*, 8(1):1–16, 2017.

[55] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William T. Freeman. Eulerian video magnification for revealing subtle changes in the world. *ACM Transactions on Graphics (Proc. SIGGRAPH 2012)*, 31(4), 2012.

[56] Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, and Nicu Sebe. Learning deep representations of appearance and motion for anomalous event detection. *arXiv preprint arXiv:1510.01553*, 2015.

[57] Chao Yan, Frans Coenen, and Bailing Zhang. Driving posture recognition by joint application of motion history image and pyramid histogram of oriented gradients. *International journal of vehicular technology*, 2014, 2014.

[58] Yichao Zhang, Silvia L Pintea, and Jan C Van Gemert. Video acceleration magnification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 529–537, 2017.