Unsupervised Disentanglement of Pose, Appearance and Background from Images and Videos

Aysegul Dundar, Kevin J. Shih, Animesh Garg, Robert Pottorf, Anrew Tao, Bryan Catanzaro

Abstract—Unsupervised landmark learning is the task of learning semantic keypoint-like representations without the use of expensive input keypoint annotations. A popular approach is to factorize an image into a pose and appearance data stream, then to reconstruct the image from the factorized components. The pose representation should capture a set of consistent and tightly localized landmarks in order to facilitate reconstruction of the input image. Ultimately, we wish for our learned landmarks to focus on the foreground object of interest. However, the reconstruction task of the *entire* image forces the model to allocate landmarks to model the background. Using a motion-based foreground assumption, this work explores the effects of factorizing the reconstruction task into separate foreground and background reconstructions in an unsupervised way, allowing the model to condition only the foreground reconstruction on the unsupervised landmarks. Our experiments demonstrate that the proposed factorization results in landmarks that are focused on the foreground object of interest when measured against ground-truth foreground masks. Furthermore, the rendered background quality is also improved as ill-suited landmarks are no longer forced to model this content. We demonstrate this improvement via improved image fidelity in a video-prediction task. Code is available at <https://github.com/NVIDIA/UnsupervisedLandmarkLearning>

Index Terms—unsupervised landmarks, keypoints, foreground-background separation, video prediction

1 INTRODUCTION

Pose prediction is a classical computer vision task that involves inferring the location and configuration of deformable objects within an image. It has applications in human activity classification, finding semantic correspondences across multiple object instances, and robot planning to name a few. One of the caveats of this task is that annotation is very expensive. Individual object “parts” need to be carefully and consistently annotated with pixel-level precision. Our work focuses on the task of *unsupervised* landmark learning, which aims to find unsupervised pose representations from image data without the need for direct pose-level annotation.

A good visual landmark should be tightly localized, consistent across multiple object instances, and grounded on the foreground object of interest. Tight localization is important because many objects (such as persons) are highly deformable. A landmark localized to a smaller, rigid area of the object will offer more precise pose information in the event of object motion. Consistency across multiple object instances is also important, as we wish for our landmarks to apply to all instances within a visual category. Finally, and most relevant to our proposed method, we want our landmarks to focus on the foreground objects. A landmark that fires on the background is a wasted landmark, as the background is constantly changing, and yields little

information regarding the pose of our foreground object of interest.

Many unsupervised landmark learning methods perturb an input training image with various transformations, then require the model to learn semantic correspondences across the transformed variants to piece together the unaltered input image. The primary issue with this approach is it penalizes the *entire* image reconstruction when we care only about the foreground, resulting in landmarks being allocated to the background. This poses a number of issues, including increased memory requirements as more landmarks required to capture the overall image and lower landmark reliability as landmarks assigned to background are unstable. Our proposed method aims to reduce the likelihood of landmarks being allocated to the background, thereby improving overall landmark quality and reducing the number of landmarks required to achieve state-of-the-art performance.

Our work builds upon existing methods in image-reconstruction-guided landmark learning techniques [8], [14], [33]. As with prior methods, we use video-based data to identify corresponding visually-consistent but spatially varied landmarks across separate frames. However, we extend this paradigm by baking in the assumption that the parts of the image we wish to model with the landmarks are *only* those that move. By adding an additional pixel-copying module to reconstruct the static image content (loosely assumed to be background), we alleviate the need for the landmarks to model the background which is ill-suited. This allows us to factorize the reconstruction task into separate foreground and background reconstructions. We further show that our method improves landmark quality even for image based datasets with fake motion synthesized by

- E-mail: {adundar, kshih, animeshg, rpottorf, atao, bcatanzaro}@nvidia.com
- A. Dundar is with Department of Computer Science, Bilkent University, Ankara, Turkey.
- A. Dundar, K.J. Shih, A. Garg, R. Pottorf, A. Tao, B. Catanzaro are with NVIDIA, Santa Clara, CA 95051.

Manuscript received xxx

using thin-plate-spline warping. In such setting, the static background assumption does not hold, and our method do not separate foreground and background. Nevertheless, we show that the foreground decoder learns to reconstruct the visually consistent features across images, and therefore learns accurate unsupervised landmarks. Our contributions are as follows:

- 1) We propose an improvement to reconstruction-guided unsupervised landmark learning that allows the landmarks to better focus on the foreground both on video and image based datasets.
- 2) We demonstrate through detailed empirical analysis that our proposed factorization allows for state-of-the-art landmark results with fewer learned landmarks, and that fewer landmarks are allocated to modeling background content.
- 3) We demonstrate that the overall quality of the reconstructed frame is improved via the factorized rendering, and include an application to the video-prediction task. We compare our video prediction results with state-of-the-art latent-representation based methods [12] and achieve better LPIPS score [40] which correlates with human perception.

2 RELATED WORKS

Our work builds upon prior methods in unsupervised discovery of image correspondences [8], [9], [14], [32], [33], [34], [41]. Most relevant here are [8] and [14], which learn the latent landmark activation maps via an image factorization and reconstruction pipeline. Each image is factored into pose and appearance representations and a decoder is trained to reconstruct the image from these latent factors. The loss is designed such that accurate image reconstruction can only be achieved when the landmarks activate at consistent locations between an image and its thin-plate-spline-warped variant. One limitation of these works is that the appearance and pose vectors also need to encode background information in order to reconstruct the entire image. To encode the background, some landmarks should fire in the background which is an ill-defined set-up because background is constantly changing and cannot be represented with consistent landmarks. Because of this limitation, previous methods [8], [14] use foreground segmentation masks when available which also requires expensive human annotation. Our work attempts to resolve this limitation by introducing unsupervised foreground-background separation into the pipeline, allowing the pose and appearance vectors to focus on modeling the foreground content.

There are few other works that propose to separate foreground and background in image rendering tasks. [15], [16] incorporates foreground-background separation with 2D keypoints annotations and applying a set of morphological operations to those keypoints such that it is able to approximately cover the whole human body. These methods are specialized for human body and cannot work for arbitrary objects. Balakrishnan et al. [2] separates foreground and background for image synthesis in an unseen pose, but their method relies on supervised 2D keypoints. Others [23], [24] separate background from foreground for single and multi-person pose-estimation with the background images

being computed by taking the median pixel value across all frames, and therefore require video sequence data with perfectly static backgrounds. Instead, our approach trains a network to synthesize a clean background from any input frame. It is therefore more forgiving with respect to background variation, and can even handle thin-plate-spline warped backgrounds. This allows our method to learn improved landmarks even on non-video datasets such as CelebA faces [13].

We also include an application to the video prediction task. Several methods have been proposed for video prediction problem, which in high level can be categorized as flow-based methods where methods model motion transformation via flows [22], [29], [37], [39] and latent-representation based methods [11], [12], [19], [31], [42]. Flow-based methods find the pixel correspondences between frames to move them accordingly in future frames. They can tackle the video-prediction problem over short-temporal ranges, but they have difficulty synthesizing previously occluded pixels. Our method lies in the latent-representation based methods. These methods encode images into latent vectors and transform latent vectors to reconstruct future frames. These methods can synthesize novel pixels, but currently they only work on simpler datasets such as KTH [27] and BAIR [6]. Most of latent-representation based methods [1], [4], [11], [12], [26] encode the background, object pose and appearance into one latent representation. There are few methods that separate pose and appearance [5], [28], [35], [38] but not background.

Recently, there have been great interest in utilizing supervised and unsupervised keypoints for video prediction [10], [28], [36], [42], and object animation tasks [29], [30], which are similar to the video prediction application we demonstrate. These work achieve maintaining structural integrity of objects in the synthesized videos thanks to the keypoint representations. The differences with our work as follows, [28] does not disentangle the background and foreground and achieves low fidelity to the initial frame. The most similar to our work is [10] which does not separate the foreground and background images, but instead learns a soft-weighting mechanism to blend the synthesized image and source image to improve static background. Their pipeline do not have a mechanism that can output a clean background and a separate foreground, and as a result the video prediction results have artifacts around the moving object, and cannot work with image-based datasets. [29] is a flow-based approach and do not have disentanglement of background whereas [30] improves upon that method with occlusion masks that are inferred from flow based dense motions. These occlusion masks are used to erase features from source images and do not provide an explicit background-foreground separation.

3 METHOD

Our method extends the pipeline proposed by [8], [14] which at a high-level reconstructs an image from two perturbed variants: one where the appearance (color, lighting, texture) information is perturbed, and one where the pose (position, orientation) of the object is perturbed. The appearance can be perturbed with color jittering (random changes

Fig. 1: This figure depicts an overview of our training pipeline on video data. Given an image frame x , we produce $T_{cj}(x)$ and $T_{temp}(x)$, which are appearance and pose perturbed variants of x respectively. The model learns to combine the appearance information from $T_{temp}(x)$, and combine it with the pose from the $T_{cj}(x)$ in order to reconstruct foreground object from the foreground decoder. Foreground masks are predicted as part of the pipeline to separate the foreground rendering from the background rendering. Specifically, the background is rendered from a UNet that learns to extract clean backgrounds from $T_{temp}(x)$. This allows the learned pose representation to focus on the more dynamic foreground object. The pose encoder and MaskNet are each depicted twice as they are applied twice during the forward pass.

to the color, brightness, and saturation), and an image of the object whose pose is perturbed can be obtained by sampling a distant frame in the same video sequence. The model must learn to extract the pose information from the appearance-perturbed image, and appearance information from the pose-perturbed image. The model will learn a set of landmarks in the process as a means to spatially-align the information extracted from the two sources in order to image reconstruction objective function. However, such objective function will penalize equally for both foreground and background pixels, and the keypoint-like landmarks are clearly not suited for the ill-defined “pose” of the background content.

Our work proposes to factorize the final reconstruction into separate foreground and background renders (see Fig. 1) with a novel design, where only the foreground is rendered conditioned on the landmark positions and appearance. The background will be inferred directly from the pose-perturbed input image with a simple UNet [25] which is depicted as *BGNet* in Fig. 1. The *BGNet* does not have the capability to model changes in pose, nor does it have access to an input image with the target pose and so we expect it to learn a simple copy function to copy background pixel content into the reconstruction for most of the pixels. In addition, we expect *BGNet* to remove the foreground object and fill the content for background. The remaining content that has gone through complex pose changes (e.g. limb motion, object rotations) will then be captured by the landmark representations. A *MaskNet* is learned that infers foreground masks for blending the foreground and background outputs for the final image reconstruction. It is applied to both input images (see Fig. 1) as foreground pose should be preserved from one input (T_{cj}) and removed

from the other (T_{temp}). Note that T_{cj} is the transformation operation that applies color jitter to the images, whereas T_{temp} is a sampling function which randomly samples a future frame with respect to image x . To achieve the static background across $T_{temp}(x)$ and x images, if cropping is applied, images are cropped from the same coordinates.

During training, this factorization is guided by a *moving foreground assumption*, in which the object of interest goes through a change in pose, while the background content remains relatively static. Nevertheless, we demonstrate that landmark quality is improved even when this assumption is held weakly. The overall architecture is trained with image reconstruction loss, and all the intermediate representations are learned end-to-end through self supervision. The various reconstruction constraints on the training pipeline force the model to disentangle pose, appearance, and background and learn foreground masks. In the next sections, we will describe the overall architecture and training pipeline in more details.

3.1 Model Components

Our full pipeline comprises five components: the pose and appearance encoders, foreground decoder, background reconstruction subnet, and mask subnet.

The pose encoder $\Phi^{pose} = Enc^{pose}(x)$ takes an input image x and returns a set of part activation maps. Critically, we want these part activation maps to be invariant to changes in local appearance, as well as to be consistent across deformations. A heatmap that activates on a person’s right hand should be invariant across varying skin tones and lighting conditions, as well track the right hand’s location across varying deformations and translations.

The *appearance encoder* $\Phi^{app} = Enc^{app}(x; Enc^{pose}(x))$ extracts local appearance information, conditioned on the pose-encoder's activation maps. Given an input image x , the pose encoder will first provide $K \times H \times W$ part activation maps Φ^{pose} . The appearance encoder then projects the image to a $C \times H \times W$ appearance feature map M^{app} . Using the pose activation map to compute a weighted sum over the appearance feature map, we extract the reduced appearance vector for the k th landmark as:

$$\Phi_{k,c}^{app} = \sum_i^H \sum_j^W \Phi_{k,i,j}^{pose} M_{c,i,j}^{app} \text{ for } c = 1 \dots C, \quad (1)$$

giving us K C -dimensional appearance vectors. Here, each activation map in Φ^{pose} is softmax-normalized.

The method pipeline attempts to reconstruct the original foreground image by combining the pose information from the K activation maps with the pooled appearance vectors for each of the K parts. Following [14], we fit a 2D Gaussian to each activation of the K activation maps by computing the mean over activation locations and using either an estimated or predefined covariance matrix. Each part is then written as: $\tilde{\Phi}_k^{pose} = (\mu_k, \Sigma_k)$, where $\mu_k \in \mathbb{R}^2$ and $\Sigma_k \in \mathbb{R}^{2 \times 2}$. The 2D Gaussian approximation forces each part activation map into a unimodal representation, introduces an information bottleneck and thereby allowing for the keypoint-like interpretation that each landmark appears in at most one location per image.

The foreground decoder (*FGDec*) and background reconstruction subnet (*BGNet*) are networks that attempt to reconstruct the foreground and background respectively. Our foreground decoder is based on the architecture proposed in SPADE [20], as was also used in [28]. SPADE is an architecture that can produce high-quality semantic-map conditioned image synthesis. Here, we use the projected Gaussian heatmaps from each 2D Gaussian landmark as a semantic maps and provide that as input to the SPADE architecture. In [28], it was shown that SPADE decoder improves the keypoint quality as the predicted keypoints input to the network in multiple scales which helps the localization of them in high resolutions.

Unlike the foreground decoder which is conditioned on bottlenecked landmark representations, the *BGNet* is given direct access to image data. Given a static background video sequence, we assume it is easier for the *BGNet* to learn to *directly copy* background pixels than it is for the landmark representations to learn to model the background. In the absence of a *BGNet*-like module, several landmarks will be allocated to capture the "pose" of the background, despite being ill-suited for such a task.

The final module is the *foreground mask subnet* (*MaskNet*), which infers the blending mask to composite the foreground and background renders. It can be interpreted as a foreground segmentation mask and is conditioned on $\tilde{\Phi}^{pose}$. *MaskNet* is also used for the background reconstruction pipeline, specifically its output is used to erase the foreground object. *BGNet* inputs the pose perturbed image with foreground object erased, then it has a simple job of copying over visible pixels and hallucinating content in locations disoccluded by the foreground object's change in pose.

In summary, the pose encoder learns explicit unsupervised landmarks due to the the input perturbations and forced unimodal 2D Gaussian representation bottleneck. The background network learns to reconstruct the background as this allows it to achieve a lower reconstruction loss.

3.2 Training Pipeline

All network modules are jointly trained in a fully self-supervised fashion, using the final image reconstruction task as guidance. We follow the training method as detailed in [14], with the addition of our proposed factorized rendering pipeline in the reconstruction phase.

Training involves reconstructing an image from its appearance and pose perturbed variants, learning to extract the unperturbed element from each variant. As with [14], we use color jittering to construct the appearance-perturbed variant $T_{cj}(x)$. When training from video data, we temporally sample a frame 3 to 60 time-steps apart from the same scene to attain the pose-perturbed variant $T_{temp}(x)$. However, in the absence of video data, we use thin-plate-spline warping to create synthetic pose-changes $T_{tps}(x)$. In general, our method is able to work with both $T_{temp}(x)$ and $T_{tps}(x)$, though TPS-warping has the downside of also warping the background pixels, making the task of *BGNet* more difficult. Let $\tilde{\Phi}^{pose}$ be the gaussian-heatmap fitted to the raw activation map Φ^{pose} , and let $\odot$ represent element-wise multiplication. Our training procedure as depicted in Fig. 1 can be expressed as follows:

$$\Phi_{cj}^{pose} = Enc^{pose}(T_{cj}(x)) \quad (2)$$

$$\Phi_{temp}^{pose} = Enc^{pose}(T_{temp}(x)) \quad (3)$$

$$\Phi^{app} = Enc^{app}(T_{temp}(x); Enc^{pose}(T_{temp}(x))) \quad (4)$$

$$\mathcal{M}_{cj} = MaskNet(\tilde{\Phi}_{cj}^{pose}) \quad (5)$$

$$\mathcal{M}_{temp} = MaskNet(\tilde{\Phi}_{temp}^{pose}) \quad (6)$$

$$\tilde{x}^{fg} = FGDec(\tilde{\Phi}_{cj}^{pose}, \Phi^{app}) \quad (7)$$

$$\tilde{x}^{bg} = BGNet((1 - \mathcal{M}_{temp}) \odot T_{temp}(x)), \quad (8)$$

$$\tilde{x} = \mathcal{M}_{cj} \odot \tilde{x}^{fg} + (1 - \mathcal{M}_{cj}) \odot \tilde{x}^{bg}. \quad (9)$$

Here, the goal is to minimize the reconstruction loss between the original input x and the reconstruction $\tilde{x}$. As can be seen, neither the shape encoder nor the appearance encoder are ever given direct access to the original image x . The pose information feeding into the foreground decoder *FGDec*($\cdot, \cdot$) is based on the color-jittered input image, where only the local appearance information is perturbed. The appearance information is captured from $T_{temp}(x)$ (or $T_{tps}(x)$), where the pose information is perturbed. Notice the shape encoder is also executed on both the pose-perturbed and color-jittered input images. This is necessary to map the localized appearance information for a particular landmark from its location in the pose-perturbed image to its unaltered position in $T_{cj}(x)$. Finally, the predicted foreground-background masks are computed for both the appearance and pose perturbed variants: $\mathcal{M}_{cj}$ and $\mathcal{M}_{temp}$ respectively. $\mathcal{M}_{cj}$ should have a foreground mask corresponding to the original foreground's pose, and is used to blend the foreground and background renders in the final step. $\mathcal{M}_{temp}$ is the foreground mask for the pose-perturbed

input image, and assists the *BGNet* in removing foreground information from its background render. Our reconstruction loss is a VGG perceptual-loss similar to the one used by [8] with pre-trained ImageNet weights.

3.3 Implementation Details

Architecture. We use the U-net architectures [25] for the pose encoder, appearance encoder, foreground mask subnet and background reconstruction subnet, complete with skip connections. The pose encoder has 4 blocks of convolutional downsampling (strided conv) modules. Each convolutional downsampling module has a convolution layer-Instance Normalization-ReLU and a downsampling layer. At each block, the number of filters doubles, starting from 64. The upsampling portion of the pose encoder has 3 blocks of convolutional upsampling, and the number of channels is halved at every block starting from 512. The appearance encoder network has one convolutional downsampling and one convolutional upsampling module. The foreground mask subnet has 3 blocks of convolutional downsampling and 3 blocks of upsampling, and the number of channels is 32 at each module. Similarly, the background reconstruction subnet has 3 blocks of convolutional downsampling and 3 blocks of upsampling. At each block, the number of filters doubles starting from 32. The image decoder has 4 convolution-ReLU-upsample modules. We first down-sample the appearance feature map by a factor of 16 in each spatial dimension. The number of output channels for each convolution-ReLU-upsampling module in the image decoder is 256, 256, 128, 64, and 3 respectively. We apply spectral normalization [17] to each convolutional layer.

For the BBC Pose dataset, we estimate the covariance from the part activation maps when fitting the Gaussians to compute $\tilde{\Phi}^{pose}$. For Human3.6M and CelebA, we use a fixed diagonal covariance of 0.08. In general, fixed diagonal covariances lead to better performance on the landmark regression task than fitted covariance, though fitted covariances lead to better image generation results. As such, we use fitted covariances for the video prediction task.

Loss Function and Optimization Parameters. We train our image factorization-reconstruction network with VGG Perceptual loss which uses the pre-trained VGG19 model provided by the PyTorch library. We apply the MSE loss on the outputs of layers `relu1_2`, `relu2_2`, `relu3_2`, and `relu4_2`, weighted by $\frac{1}{32}$, $\frac{1}{16}$, $\frac{1}{8}$, and $\frac{1}{4}$ respectively. We use the Adam optimizer with a learning rate of $1e^{-4}$, and weight decay of $5e^{-6}$. The network is trained on 8 GPUs with a batch size of 16 images per GPU.

4 EXPERIMENTS

Here, we analyze the effect of introducing foreground-background separation into an unsupervised-landmark pipeline. Through empirical analysis, we demonstrate that the learned landmarks are used for capturing foreground information, thereby improving overall landmark quality. Landmark quality is evaluated by using linear regression to map the unsupervised landmarks to annotated keypoints, with the assumption that well-placed, spatially consistent landmarks lead to low regression error. Finally, we include

an additional application of our method in the video prediction task, demonstrating how the factorized rendering pipeline improves the overall rendered result.

4.1 Datasets

We evaluate our method on Human3.6M [7], BBC Pose [3], CelebA [13], KTH [27], and BAIR action-free [6] datasets. Human3.6M is a video dataset that features human activities recorded with stationary cameras from multiple viewpoints. BBC Pose dataset contains video sequences featuring 9 unique sign language interpreters. Individual frames are annotated with keypoint annotations for the signer. While most of the motion is from the hand gestures of the signers, the background features a constantly changing display that makes clean background separation more difficult. CelebA is an image-only dataset that features keypoint-annotated celebrity faces. As with prior works, we separate out the smaller MAFL subset of the dataset, train our landmark representation on the remaining CelebA training set, and perform the annotated regression task on the MAFL subset. The KTH dataset comprises videos of people performing one of six actions (walking, running, jogging, boxing, hand-waving, hand-clapping). BAIR dataset consists of videos with robot arms moving randomly with a diverse set of objects on a table. We use KTH and BAIR datasets for our video prediction application.

Dataset Preprocessing. For BBC Pose, we first roughly crop around each signer by using the given keypoints. Specifically, we find the center of the keypoints and crop a box of 300×300 around the center, then resize the crops to 128×128 . For the Human3.6M dataset, we follow the procedure defined by [41] for training/validation splits. We find the center of the keypoints and crop 300×300 around the center, and again, resize the crops to 128×128 . For the CelebA/MAFL dataset, we follow [8] by resizing the images to 160×160 , and center crop by 128×128 .

The thin plate spline transformation $T_{tps}(x)$ allows us to perform a non-rigid warping of the image content based on applying perturbations to a grid of control points in the image's coordinate space. It is more expressive than a standard affine transformation on the image, allowing us to deform the image content in more interesting ways, and thus a reasonable drop-in replacement for $T_{temp}(x)$ when temporal data is not available. This creates synthetic "motion," but has the downside of also warping the background. This issues is occasionally ameliorated by flat-textureless regions in the background, where warping would make little to no visual difference.

4.2 Unsupervised Landmark Evaluation

As with prior works [8], [34], we fit a linear regressor (without intercept) to our learned landmark locations from our pose representation to supervised keypoint coordinates. Following [8], we create a loose crop around the foreground object using the provided keypoint annotations, and evaluate our landmark learning method within said crop. For a fair comparison, we allow the number of learnable landmarks same as the previous methods. Specifically, we use 30 landmarks for the BBC Pose dataset, 16 landmarks for Human3.6M, and 10 landmarks for CelebA dataset.

TABLE 1: Evaluation of landmark accuracy on Human3.6M, BBC Pose, and MAFL. Human3.6M error is normalized by image dimensions. For BBC Pose, we report the percentage of annotated keypoints predicted within a 6-pixel radius of the ground truth. For MAFL, prediction error is scaled by inter-ocular distance.

Human3.6M		Error
supervised	Newell et al. [18]	2.16
unsupervised	Thewlis et al. [33]	7.51
	Zhang et al. [40]	4.91
	Lorenz et al. [14]	2.79
	Baseline (temp)	3.07
	Baseline (temp,tps)	2.86
	Ours	2.73
(a)		
BBC Pose		Acc.
supervised	Charles et al. [3]	79.9%
	Pfister et al. [21]	88.0%
unsupervised	Jakab et al. [8]	68.4%
	Lorenz et al. [14]	74.5%
	Baseline (temp)	73.3%
	Baseline (temp,tps)	73.4%
	Ours	78.8%
(b)		
MAFL		Error
unsupervised	Thewlis et al. [33]	6.32
	Zhang et al. [40]	3.46
	Lorenz et al. [14]	3.24
	Jakab et al. [8]	3.19
	Baseline (tps)	4.34
	Median Filtering	4.33
	Ours (No Mask)	2.88
	Ours	2.76
(c)		

First, we report our regression accuracies on video datasets, Human3.6M and BBC Pose datasets. For the video datasets, we found it best to use only $T_{temp}(x)$ to sample perturbed poses from future frames during training. Our primary baseline is our model without the explicit foreground-background separation. For this baseline, we report results using T_{temp} -only (Baseline (temp)) as well as both T_{temp} and T_{tps} (Baseline (temp, tps)). Results are shown in Tables 1a, 1b. In all cases, we demonstrate that including factorized foreground-background rendering improves landmark quality compared to the controlled baseline model. Our method also beats the competing methods significantly.

Next, we analyze how factorizing out the background rendering influences landmark quality. In Fig. 2a, we present an ablation study where we measure the regression-to-annotation accuracy against the number of learned landmarks. Compared to our baseline models, we see that the background-factorization allows us to achieve better accuracy with fewer landmarks, and that the degradation is less steep. In Fig. 2b, we evaluate our baselines and our method by translating the unsupervised keypoints to supervised ones with linear regressors trained with different number of supervised examples. Supervised examples are chosen by a random sampler and experiments are repeated 5 times with different random seeds. Mean accuracies and standard deviations are reported in Fig. 2b. Whereas the results are comparable when there are only 1 or 10 supervised examples available, our method starts to show significant improvements after 100 supervised examples,

(a) BBC validation dataset keypoint accuracy versus number of learned keypoints. By factorizing out the background rendering, we are able to achieve better landmark-to-annotation mappings with fewer landmarks than the baseline.

(b) Performance of landmark detector trained on BBC dataset as a function of the number of supervised examples that are used to translate unsupervised landmarks to supervised ones.

(c) Percentage of the per-landmark normalized activation maps contained within the provided foreground segmentation masks on Human3.6M, sorted in ascending order. We compare our model against our baseline at 8, 12, and 16 learned landmarks. We see that the least-contained landmarks in the proposed approach are significantly more contained than those of the baseline.

Fig. 2: Landmark analysis experiments.

and the improvements are preserved as the number of supervised examples increases. Next, we validate one of our primary claims which is that by factorizing foreground and background rendering in the training pipeline, we allow the landmarks to focus on modeling the pose and appearance of the foreground objects, leaving the background rendering task to a less expressive, but easier to learn mechanism. Human3.6M dataset provides foreground-background segmentation masks. If the landmarks truly focus more on modeling the foreground more, then underlying activation heatmaps for each unsupervised landmark should be more contained within the provided segmentation masks in the factorized case. In Fig. 2c, we compare the percentage of the normalized activation maps contained within the provided segmentation masks against our baseline model for 8, 12, and 16 landmark models. Specifically, for the activation map of each landmark, we sum over the activations that lie in the foreground region defined by segmentation mask and

Fig. 3: Qualitative results of our landmark prediction pipeline. From top to bottom, we show our regressed annotated keypoint predictions, our predicted foreground mask, and the underlying landmark activation heatmaps. Datasets are BBC Pose, Human3.6M, and CelebA/MAFL respectively.

Fig. 4: Additional qualitative results for regressed annotated keypoint predictions. Rows show test image results for BBC Pose, Human3.6M, and CelebA/MAFL respectively.

divide the sum by the overall magnitude of the activation values. We then sort the landmarks in ascending order of containment (horizontal axis of Fig. 2c) and plot the models' landmark-containment curve. The results in Fig. 2c demonstrate that the foreground-background factorization noticeably improves the least containment of the least-contained landmarks. They are an order of magnitude larger than the corresponding baseline landmarks. It is safe to say that the least-contained landmarks for the baseline model are nearly completely utilized for modeling the image background (99%+ of the activation mass is on the background). In all of our experiments so far, we roughly cropped around the foreground object following the previous works. However, our method has the advantage of modeling the background and so the entire image can be given to it. We experiment this on Human3.6M by feeding the full image (1000×1000) to our baseline and proposed methods during training and inference. In this setting, our proposed method achieves 1.72 error, whereas baseline (temp) and baseline (temp, tps) achieve 2.66 and 2.16 error, respectively. Note that these results are not comparable with Table 1a, because training and testing images are larger and errors are normalized with image sizes. Additionally, the network architecture is optimized to handle 128×128 images. Nevertheless, the

Fig. 5: From left to right input image, thin-plate-spline warped image, reconstructed background, predicted foreground, mask, and reconstructed output. The first three rows belong to our method and the last row shows results of the median filtering experiment.

relative improvements over baseline can be analyzed within these two set-ups. As reported in Table 1a, our method improves the best performing baseline by 4.5% when the images are loosely cropped around the subject. On the other hand, when the full images are used, our method improves the baseline by 20.4%. This experiment additionally shows the benefit of our method, it does not require loose bounding boxes around objects which will not be available in practice for unsupervised pose estimation.

Lastly, we test our method on CelebA/MAFL dataset. This set-up is interesting because it is an image only dataset, and our architecture is built on motion-based foreground assumption. However, as can be seen in Table 1c, our method significantly improves the baseline even given very weak static background assumptions. This is because $T_{tps}(x)$ indiscriminately warps the entire image, creating a pose-perturbed variant with a heavily deformed background. The *BGNet* in this setting learns to rectify TPS

Fig. 6: We base our main evaluation to LPIPS score which closely correlates with human perception. We also provide SSIM and PSNR metrics for completeness. Our method achieves significantly better LPIPS score than state-of-the-art methods and competitive SSIM and PSNR scores on both KTH and BAIR datasets. It also shows a large improvement over our controlled baseline.

warped image in the training set, and as the landmark-conditioned foreground decoder is still better suited to tracking visually-consistent (backgrounds are not consistent across the dataset) facial features across warps, we are still able to achieve a weak foreground-background separation that contributes to improved landmark accuracy. This demonstrates the flexibility of our method, as a median-based filter to extract a perfectly static background would not be feasible in this scenario. We test the median-based filtering approach as well where the background image is constructed by taking median over the frames in the batch. The disentanglement result of the median filtering experiment can be seen in Fig. 5-last row. It outputs a noisy face image with background canceled out and leaves the work of reconstructing the whole image to the foreground decoder. As reported in Table 1c, median filtering approach results in the same accuracy as the baseline as both methods rely on the same foreground decoder to reconstruct the image. Further, in Table 1c, we include a No Mask baseline which is our proposed model but sans predicted blending masks. Here, we combine foreground and background directly with: $\tilde{x} = \tilde{x}^{fg} + \tilde{x}^{bg}$. This variant also improves over the unfactorized baseline, though the full pipeline still performs best.

We show qualitative results of our regressed annotated keypoint predictions, as well as landmark activation and foreground mask visualizations in Fig. 3. From top to bottom, we show our regressed annotated keypoint predictions, our predicted foreground mask, and the underlying landmark activation heatmaps. Datasets are BBC Pose, Human3.6M, and CelebA/MAFL respectively. Notice that the degree of binarization in the predicted mask is indicative of the strength of the static background assumption on the data. Human3.6M features a strongly static background, whereas BBC Pose has a constantly updating display on the

left, and CelebA was trained with T_{tps} which indiscriminately warps both foreground and background. Nevertheless, our method still improves over the baseline despite imperfectly binarized foreground-background separation. Fig. 4 shows additional results for the pose-regression task on various datasets. The regression quality is generally very accurate for various actors in wide range of motions.

Further Discussion on Landmark Evaluation Results.

Our architecture is built on a motion-based foreground assumption. However, we still see a significant improvement in the landmark quality in the CelebA dataset for reasons explained in this section.

Our architecture includes two potential streams that can reconstruct a given image in static datasets, one is the appearance and pose encoding-decoding branch that uses a Gaussian-landmark bottlenecked representation, and the other is the background network comprising a small UNet directly mapping input RGB data to the output reconstruction. The *BGNet* does not have the ability to adapt its behavior to changes in pose the way the landmark-conditioned foreground decoder can. As such, it cannot perfectly reconstruct the training images from novel deformations. On the other hand, the landmark-conditioned foreground decoder is better suited to tracking visually-consistent (backgrounds are not consistent across the dataset) facial features across warps. Therefore, the training process is more likely to assign the foreground decoder to reconstruct the visually consistent features to minimize the loss. We test this hypothesis by training our foreground decoder, background decoder, and the overall pipeline separately to reconstruct the unaltered images. In this setting, the background decoder given a TPS warped image should learn to rectify it. We measure the reconstruction accuracy with the VGG Perceptual loss between the reconstructed images and original inputs on CelebA

validation set. First, we measure the VGG Perceptual loss between TPS warped input images and unaltered images as 2.63 providing us with an upper bound. When the image is only reconstructed with the foreground decoder, this loss goes down to 1.00, it is still high because of the bottleneck in the pose-appearance disentanglement. When only the background decoder is trained to reverse the TPS warping, the loss comes as 0.56. The background decoder is not provided with the original part based locations, neither provided with the random TPS variables, still it learns to rectify the image using clues from the input. These clues may be the artefacts caused by the TPS warping, which may be helping the network to estimate how to rectify it. When we reconstruct the images with the overall pipeline, the reconstruction loss goes down to 0.48. In the end, to achieve this low reconstruction loss, the network decides to utilize both the foreground and background networks and achieves a weak foreground-background separation that contributes to improved landmark accuracy.

In Fig. 5, we show our model outputs from training on CelebA. We see that the pipeline has determined the parts that contain eyes, nose, mouth, and eyebrows (consistent parts among examples) to be foreground, with everything else as background. Importantly, the fourth column from the left shows that the *BGNet* is capable of learning how to rectify a TPS warped image. Because there is no true moving foreground in this dataset, *BGNet* does not need to generate novel pixels for the previously occluded pixels, nor does it need to remove the face as it will appear in the same place and can be handled by the foreground mask.

4.3 Application to Video Prediction

Lorenz et. al [14] applied their model to video-to-video style transfer on videos of BBC signers, indicating that the rendered images from the landmark model are temporally stable and [28] extended this work to the video prediction task. One of the issues with these renders, however, is that the landmarks are not suited for modeling the background, resulting in low-fidelity rendered backgrounds. We demonstrate that our factorized formulation better handles this issue.

We evaluate our rendering on the video prediction task on the KTH and BAIR datasets, and compare against external methods. Following the implementation in [14], [28], we assume the appearance information remains constant throughout each video sequence, and use a deterministic LSTM to predict how the 2D Gaussians move through time conditioned on an initial set of seed-frames. We follow the same procedure of competing methods and for KTH dataset, condition LSTMs on 10 initial frames to predict the next 30 frames and for BAIR dataset, condition LSTMs on 2 initial frames to predict the next 28.

LSTM training details. The LSTM comprises 3 LSTM layers and a final linear layer. Each LSTM layer has 256 channels. For the KTH results, we trained our landmark model with 40 landmarks. The LSTM was trained to predict 10 future frames from an input sequence of 10. For the BAIR dataset, we trained our landmark model with 30 landmarks, and the LSTM is trained on an input sequence of length 10. On the KTH video prediction task, we modify the image

decoder's first convolutional block to run at both 1/8 and 1/16 scales, and elementwise sum the resulting tensors together after 2x upscaling the latter. This is done for both the proposed and baseline model, as we found running the decoder from a 1/16th downsampled tensor to be too aggressive for this dataset at 64×64 .

We show our quantitative results in Fig. 6. We base our main evaluation to LPIPS score [40] which closely correlates with human perception and also report SSIM and PSNR for completeness. Note that the background-factorized approach significantly outperforms the unfactorized baseline on all performance metrics, indicating better background reconstruction, as the foreground is a comparatively smaller portion of the frame. Our method also achieves significantly better LPIPS score than the competing methods. While the other methods lose accuracy faster as the predicted frame moves more into the future, baseline and our method can achieve longer-range video prediction. Also note that, SAVP deterministic model even though outputs blurry images and achieves poor LPIPS scores, it has the best SSIM/PSNR scores on BAIR datasets, whereas our method outputs sharp images which results in best LPIPS scores as well as competitive SSIM and PSNR scores on both datasets.

In Fig. 7 and 8, we show our estimated landmarks (only few of them are displayed), rendered foregrounds conditioned on estimated landmarks and appearance code, rendered masks conditioned on landmarks, rendered backgrounds, and the corresponding compositions. Our method assumes a fixed background for the entire sequence, but predicts a new foreground and blending mask for each extrapolated timestep. Both our baseline and proposed method maintain better structural integrity than other methods. However, due to the imperfect binarization of the predicted mask, the foreground in the composite image may appear somewhat faded compared to that of other methods. Improved binarization of the predicted masks remains a topic of future work. In Fig. 7, DRNET [5], SAVP deterministic [12], and SAVP stochastic [12] methods produce blurry images where person loses structural integrity whereas our method produces sharp foregrounds. As shown in Fig. 8, SVGLP [4] and SAVP stochastic [12] output artifacts at the previously occluded regions, whereas SAVP deterministic generates blurry robot arm. Our method outputs sharp images with less artifacts.

5 CONCLUSION

We propose and study the effects of explicitly factorized foreground and background rendering on reconstruction-guided unsupervised landmark learning. Our experiments demonstrate that by careful architectural design, we can disentangle an image into pose, appearance, and background, and learn foreground masks. With this setup, the model can do a better job of allocating landmarks to the foreground objects of interest. As such, we are able to achieve more accurate regressions to annotated keypoints with fewer landmarks, thereby reducing memory requirements. The disentangled representations can also be useful for image manipulation applications as can be seen in our unsupervised-landmark-based video prediction task.

Fig. 7: Qualitative results on KTH action test dataset comparing our method to prior work. Our baseline produces a sharp foreground, but the background does not match that of the initial frames. Our proposed factorized rendering significantly improves the background fidelity. The bottom four rows show our factorized outputs. From top to bottom, we have estimated landmarks representation via LSTM, the rendered foreground conditioned on estimated landmarks (parameterized as 2DGaussian) and appearance vector encoded from the first frame, the predicted blending mask conditioned on estimated landmarks, and the rendered background (first image on bottom row) followed by the composite output.

REFERENCES

- [1] M. Babaeizadeh, C. Finn, D. Erhan, R. H. Campbell, and S. Levine. Stochastic variational video prediction. *arXiv preprint arXiv:1710.11252*, 2017.
- [2] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. Guttag. Synthesizing images of humans in unseen poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8340–8348, 2018.
- [3] J. Charles, T. Pfister, D. Magee, D. Hogg, and A. Zisserman. Domain adaptation for upper body pose tracking in signed TV broadcasts. In *British Machine Vision Conference*, 2013.
- [4] E. Denton and R. Fergus. Stochastic video generation with a learned prior. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.
- [5] E. L. Denton et al. Unsupervised learning of disentangled representations from video. In *Advances in neural information processing systems*, pages 4414–4423, 2017.
- [6] C. Finn, I. Goodfellow, and S. Levine. Unsupervised learning for physical interaction through video prediction. In *Advances in neural information processing systems*, pages 64–72, 2016.
- [7] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013.
- [8] T. Jakab, A. Gupta, H. Bilen, and A. Vedaldi. Unsupervised learning of object landmarks through conditional image generation. In *Advances in Neural Information Processing Systems*, 2018.
- [9] A. Kanazawa, D. W. Jacobs, and M. Chandraker. WarpNet: Weakly supervised matching for single-view reconstruction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [10] Y. Kim, S. Nam, I. Cho, and S. J. Kim. Unsupervised keypoint learning for guiding class-conditional video prediction. In *Advances in Neural Information Processing Systems*, pages 3809–3819, 2019.
- [11] M. Kumar, M. Babaeizadeh, D. Erhan, C. Finn, S. Levine, L. Dinh, and D. Kingma. Videoflow: A flow-based generative model for video. *arXiv preprint arXiv:1903.01434*, 2019.
- [12] A. X. Lee, R. Zhang, F. Ebert, P. Abbeel, C. Finn, and S. Levine. Stochastic adversarial video prediction. *arXiv preprint arXiv:1804.01523*, 2018.
- [13] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.
- [14] D. Lorenz, L. Bereska, T. Milbich, and B. Ommer. Unsupervised part-based disentangling of object shape and appearance. In *CVPR*, 2019.
- [15] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool. Pose guided person image generation. In *Advances in Neural*

Fig. 8: Qualitative results on the BAIR dataset. SAVP deterministic model outputs blurry robot arm. SVGLP and SAVP (stochastic) models outputs artifacts at the pixels previously occluded by the robot arm. Baseline model is not able to reconstruct the objects perfectly. Please zoom into see each object is slightly different than the corresponding ground-truth object. Our method separates robot arm and background image, outputs a realistic video.

- Information Processing Systems, pages 406–416, 2017.
- [16] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz. Disentangled person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 99–108, 2018.
- [17] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018.
- [18] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016.
- [19] J. Oh, X. Guo, H. Lee, R. L. Lewis, and S. Singh. Action-conditional video prediction using deep networks in atari games. In *Advances in neural information processing systems*, pages 2863–2871, 2015.
- [20] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019.
- [21] T. Pfister, J. Charles, and A. Zisserman. Flowing convnets for human pose estimation in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1913–1921, 2015.
- [22] F. A. Reda, D. Sun, A. Dundar, M. Shoeybi, G. Liu, K. J. Shih, A. Tao, J. Kautz, and B. Catanzaro. Unsupervised video interpolation using cycle consistency. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 892–900, 2019.
- [23] H. Rhodin, V. Constantin, I. Katircioglu, M. Salzmann, and P. Fua. Neural scene decomposition for multi-person motion capture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7703–7713, 2019.
- [24] H. Rhodin, M. Salzmann, and P. Fua. Unsupervised geometry-aware representation for 3d human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 750–767, 2018.
- [25] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 2015.
- [26] M. Saito, E. Matsumoto, and S. Saito. Temporal generative adversarial nets with singular value clipping. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2830–2839, 2017.
- [27] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 32–36. IEEE, 2004.
- [28] K. J. Shih, A. Dundar, A. Garg, R. Pottorf, A. Tao, and B. Catanzaro. Video interpolation and prediction with unsupervised landmarks.

- arXiv preprint arXiv:1909.02749*, 2019.
- [29] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2377–2386, 2019.
 - [30] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe. First order motion model for image animation. In *Advances in Neural Information Processing Systems*, pages 7135–7145, 2019.
 - [31] N. Srivastava, E. Mansimov, and R. Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852, 2015.
 - [32] S. Suwajanakorn, N. Snavely, J. J. Tompson, and M. Norouzi. Discovery of latent 3d keypoints via end-to-end geometric reasoning. In *Advances in Neural Information Processing Systems*, pages 2059–2070, 2018.
 - [33] J. Thewlis, H. Bilen, and A. Vedaldi. Unsupervised learning of object frames by dense equivariant image labelling. In *Advances in Neural Information Processing Systems*, pages 844–855, 2017.
 - [34] J. Thewlis, H. Bilen, and A. Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *International Conference on Computer Vision (ICCV)*, 2017.
 - [35] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1526–1535, 2018.
 - [36] R. Villegas, J. Yang, Y. Zou, S. Sohn, X. Lin, and H. Lee. Learning to generate long-term future via hierarchical prediction. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3560–3569. JMLR. org, 2017.
 - [37] J. Walker, A. Gupta, and M. Hebert. Dense optical flow prediction from a static image. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.
 - [38] N. Wichers, R. Villegas, D. Erhan, and H. Lee. Hierarchical long-term video prediction without supervision. *arXiv preprint arXiv:1806.04768*, 2018.
 - [39] T. Xue, J. Wu, K. Bouman, and B. Freeman. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In *Advances in Neural Information Processing Systems*, 2016.
 - [40] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.
 - [41] Y. Zhang, Y. Guo, Y. Jin, Y. Luo, Z. He, and H. Lee. Unsupervised discovery of object landmarks as structural representations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
 - [42] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. Metaxas. Learning to forecast and refine residual motion for image-to-video generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 387–403, 2018.

Aysegul Dundar Aysegul Dundar is an Assistant Professor of Computer Science at Bilkent University, Turkey and a Sr. Research Scientist at NVIDIA. She received her Ph.D. degree at Purdue University in 2016, under supervision of Professor Eugenio Culurciello. She received a B.Sc. degree in Electrical and Electronics Engineering from Bogazici University in Turkey, in 2011. In CVPR 2018, she won the 1st place in the Domain Adaptation for Semantic Segmentation Competition in the Workshop on Autonomous Vehicle challenge. Her current research focuses are on domain adaptation, image segmentation, and generative models for image synthesis and manipulation.

Kevin J. Shih Kevin J. Shih is a Research Scientist in the Applied Deep Learning Research team at NVIDIA. He obtained his Ph.D. in Computer Science from University of Illinois at Urbana–Champaign under the supervision of Professor Derek Hoiem. Prior to that, he received his B.S.E from the University of Michigan. His research interests include object localization, pose estimation, attention mechanisms, and models that handle multiple modalities.

Animesh Garg Animesh Garg is an CIFAR AI Chair Assistant Professor of Computer Science at University of Toronto, a Faculty Member at the Vector Institute, and a Sr. Research Scientist at NVIDIA. Animesh earned his PhD from UC Berkeley and completed a postdoc at Stanford. His current research focuses on machine learning algorithms for perception and control in robotics. His interests combine reinforcement learning with computer vision for applications in mobile-manipulation and surgical robotics.

Robert Pottorff Robert Pottorff is a Research Scientist in the Applied Deep Learning Research team at NVIDIA. He received his masters degree in Computer Science from Brigham Young University in 2019. His research interests include dynamical system models of computer vision and human perceptual biases in video and motion.

Andrew Tao Andrew Tao is a Distinguished Engineer and Manager of the Computer Vision side of the Applied Deep Learning Research group at NVIDIA. He received his Masters in Electrical Engineering from Stanford University in 1992 with an emphasis on Computer Architecture. He has worked as a CPU hardware engineer, as GPU hardware engineer and architect, as the Director of Applied Architecture at NVIDIA, and has led a number of Computer Vision teams in the Automotive sector.

Bryan Catanzaro Bryan Catanzaro is Vice President of Applied Deep Learning Research at NVIDIA. After receiving his Ph.D. from UC Berkeley in 2011, he worked as a research scientist at NVIDIA on programming models and applications for GPUs, focusing on libraries for neural networks, which led to the creation of the CUDNN library. He worked at Baidu Silicon Valley AI Lab from 2014-2016, contributing to the DeepSpeech project. In 2016, he returned to NVIDIA to build a lab applying deep learning to problems in computer vision, graphics, speech, language, and chip design.