MetaPix: Few-Shot Video Retargeting

Jessica Lee **Deva Ramanan** **Rohit Girdhar**
Carnegie Mellon University
https://imjal.github.io/MetaPix/

Abstract

We address the task of unsupervised retargeting of human actions from one video to another. We consider the challenging setting where only a few frames of the target is available, including the case of a single target frame (e.g., video-to-*image* retargeting). The core of our approach is a conditional generative 'Pose2Im' model that can transcode input skeletal poses (automatically extracted with an off-the-shelf pose estimator) to output target frames. However, it is challenging to build a universal transcoder because humans can appear wildly different due to clothing and background scene geometry. Instead, we learn to adapt – or *personalize* – a universal generator to the particular human and background in the target. To do so, we make use of *meta*-learning to discover effective strategies for on-the-fly personalization. One significant benefit of meta-learning is that the personalized Pose2Im transcoder naturally enforces temporal coherence across its generated frames; all frames will contain consistent clothing and background geometry of the target. We experiment on in-the-wild internet videos and images and show our approach improves over widely-used baselines for the task.

1 Introduction

One of the hallmarks of human intelligence is the ability to imagine. For example, given an image of a never-before-seen person, one can easily imagine them performing different actions. To do so, we make use of years of experience watching humans act and interact with the world. Crucially, we effortlessly adapt or *retarget* universal rules of action to a specific human and environment. Our goal in this work is to develop models that learn to generate human motions by specializing universal knowledge to a particular target human and environment, given only a few samples of the target.

It is attractive to tackle such video generation tasks using the framework of generative adversarial neural networks (GANs). Past work has cast the core computational problem as one of conditional image generation where input source poses (automatically extracted with an off-the-shelf pose estimator) are transcoded into image frames [1, 5, 7]. A particularly successful approach to the pose to image generation problem is training of specialized – or *personalized* – models to particular scenes. These often require large-scale target datasets, such as 20 minutes of footage in a target lab setting [2].

The above approaches make use of personalization as an implicit but crucial ingredient, by *on-the-fly* training of a generative model tuned to the particular target domain of interest. Often, personalization is operationalized by fine-tuning a generic model on the specific target frames of interest. Our key insight is recasting personalization as an *explicit* component of a video-retargeting engine, allowing us to make use of meta-learning to *learn* how best to fine-tune (or personalize) a generic model to a particular target domain. We demonstrate that (meta)learning-to-fine-tune is particularly effective in the few-shot regime, where few target frames are available. From a technical perspective, one of our contributions is extending meta-learning to GANs, which is nontrivial because both a generator and discriminator need to be adversarially fine-tuned. To that end, we propose MetaPix, a novel approach to personalization for video retargeting.

3rd Workshop on Meta-Learning at NeurIPS 2019, Vancouver, Canada.

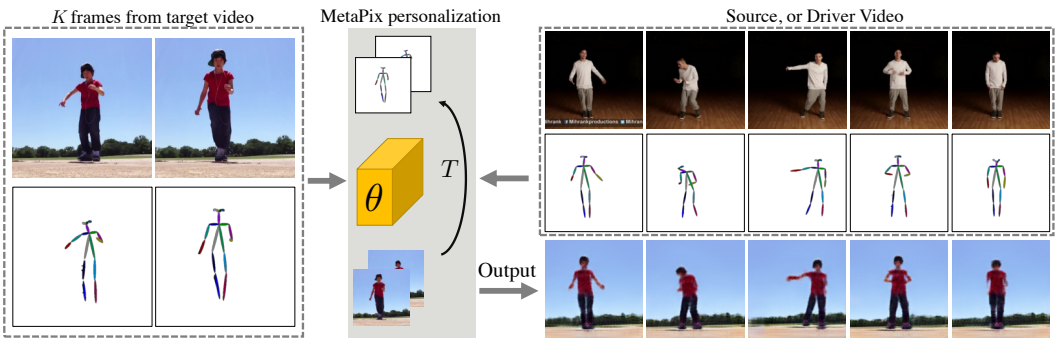

Figure 1: **Video retargeting on a budget.** Our goal is to retarget a source video to a target, *quickly* (running a few, T, iterations for adaptation) and *efficiently* (given a few, K, frames from the target domain). We achieve that by (meta)learning a model θ that is able to quickly and efficiently adapt to a given target video.

2 Our Approach

We now describe MetaPix in detail. To reiterate, our goal is to learn a generic model of human motion, parameterized by θ, that can quickly and efficiently be *personalized* for a specific person. We define speed and efficiency requirements in terms of two parameters: computation/iterations (T) and the number samples required for personalization (K), respectively.

Base retargeting architecture. We build upon popular video retargeting architectures. Notably, there are two common approaches in literature:1) Learning a transformation from one image to another, conditioned on the pose [1, 9] and 2) Learning a mapping from pose to RGB (Pose2Im), like [2]. Both obtain strong performance and amenable to the speed and efficiency constraints we are interested in. For example in K-shot setting (i.e. to learn a model using K frames), one can train the Pose2Im mapping using the K frames in the former case, or use the C_2^K pairs from K frames to learn a transformation function in the latter case. They are also both compatible with our MetaPix optimization discussed next, though for brevity and ease of implementation, we stick with the former, Pose2Im models for the rest of the paper, specifically Pix2PixHD [8] due to its strong performance.

MetaPix. MetaPix builds upon the base retargeting architecture by optimizing it for few-shot and fast adaptation for personalization. We achieve that by taking inspiration from the literature on few-shot learning, where meta-learning has shown promising results. We use a recently introduced first-order meta-learning technique, Reptile [6]. As compared to the more popular technique, MAML [3], it is more efficient as it does not compute a second gradient and is amenable to work with arbitrary optimizers as it does not need to backpropagate through the optimization process. Given that GAN architectures are hard to optimize, Reptile suits our purposes of its ability to use Adam [4], the default optimizer for Pix2PixHD, as our task optimizer.

We start with the base model trained for the Pose2Im retargeting task. We then meta-train this model by performing the following steps, further outlined in Algorithm 1 in the Appendix. In each meta-iteration, we sample a *task*: in our case a set of K frames from a new video to personalize to. We then finetune the current model parameter to that video over T iterations, and update the model parameters in the direction of the personalized parameters using a meta learning rate ϵ. Note that our base model is based on a GAN, so has two network weights to be optimized, the generator (θ_G) and discriminator (θ_D). For simplicity, we optimize both θ_D and θ_G jointly at each step, and find that performs well in practice. Implementation details are included in the appendix.

3 Experiments

3.1 Datasets and Evaluation

We meta-train our approach on a dataset of 10 dance internet videos from YouTube, where you can view sample frames in the Appendix. Due to a lack of standard benchmark for video retargeting tasks, we evaluate our method using the dataset described in [9]. We split each of the 8 test videos into a training and test sequence in 0.85:0.15 ratio, and sample K training and 2000 test frames from

Method	Init	K	T	SSIM	PSNR	MSE
Pix2PixHD	Random	∞	∞	0.68	19.56	2,427.18
Pix2PixHD	Pretrain	∞	∞	0.69	19.31	2,673.40
Pix2PixHD	Random	5	20	0.08	9.51	7,801.19
Pix2PixHD	Pretrain	5	20	0.35	12.00	5,576.28
Ours	MetaPix	5	20	0.39	13.73	4,696.00

Table 1: **Adding MetaPix.** We compare Random, Pose2Im, and MetaPix model's performance. Constraining the amount of frames and compute used at test time leads to a drop in performance as expected. However, if the model is pretrained using MetaPix, it obtains better performance. We compare these methods qualitatively in the Appendix, in Figure 6.

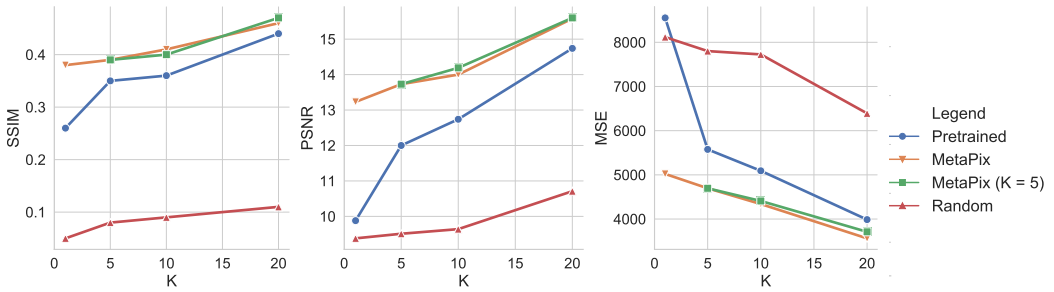

Figure 2: **Personalization using K frames.** We find that while all initializations get better with increasing K, and using MetaPix consistently outperforms simple pretraining. Moreover we note that even using a model trained with MetaPix for $K = 5$ works well at any K value used at test time, showing the generalizability of MetaPix. It is worth noting that the biggest gap is seen at lower values of K, showing our method is most useful in cases where one has little data for personalization.

the test sequence. We use the same metrics as in [9] for ease of comparison: Mean Squared Error (MSE), Structured Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR). Each of these are averaged over the 2000 test frames from each of the 8 test videos. To compare our baselines and our method, pose retargeting as a task aims to minimize MSE and maximize SSIM and PSNR.

3.2 Evaluating MetaPix

We build our baseline model based on Pix2PixHD [2, 8]. To get an upper bound on performance on this model, we train the model for each test video with no constraints on T or K, and report the performance in Table 1. We use all of the frames from first 85% of the video, and train it to convergence, which we empirically found at 10 epochs. Then, we evaluate the performance of our model in constrained settings, where we want to learn to personalize given a few samples and in a constrained computational budget. Hence, we use a pretrained model on train set and a random model, and we personalize them by finetuning on each test video. As Table 1 shows, applying constraints leads to a drop in performance in all methods, as expected from using only 5 frames finetuned over 20 iterations. Finally, we compare that to the MetaPix model: in that case, we start from the pre-trained model, and do meta-learning on top of those parameters to optimize them for the transfer task as described in Section 2. That leads to a significant improvement over the pretrained model, showing the strength of MetaPix for this task. In the Appendix in Figure 6, we visualize the predictions between the baselines and our method. It is interesting to note that the meta-learned model is able to adapt to the color of the clothing and the background much better than a pretrained model, given the same frames for personalization. This reinforces MetaPix is a much better initialization for few-shot personalization than directly finetuning from a generic pretrained model.

3.3 Ablations

We now ablate the key design choices in our MetaPix formulation. One of the strengths of our formulation is the explicit control on the supervision provided and computation the model is allowed to perform, and depending on the use-case, those parameters can easily be tweaked.

Variation in K: We vary the amount of supervision for personalization, K, and evaluate its effect on the metrics in Figure 2. We compare the following models: a) Randomly initialized, b) Pretrained

Figure 3: **Better temporal coherence with MetaPix.** *Randomly* picked frames two videos generated using pre-training and MetaPix. MetaPix leads to much more temporally coherent results with consistent clothing and backgrounds. Full video in the supplementary.

on the train set, c) Trained using MetaPix for each value of K and tested with the same K, and d) Trained using MetaPix for $K = 5$ and tested at each value of K. The last one tests the generalizability of MetaPix to different values of K at train and test time. We find that the MetaPix trained models consistently perform better than a simple pretrained model on all metrics. Notably, the model only trained for $K = 5$ is still able to obtain strong performance at different K values, showing the MetaPix trained model can generalize beyond the specific setup it is optimized for. The gap between the MetaPix trained model and the pretrained model tends to reduce with higher K, which is as expected: more data for personalization would likely reduce the importance of the initialization. However, there is a clear and significant gap for lower values of K, showing that MetaPix is highly effective for retargeting from few samples. In fact, we find that meta-learning is most effective for $K = 1$, corresponding to the challenging scenario of video-to-image retargeting.

Variation in T: Similar to variation in supervision, we experiment with varying the computation, or T, in Figure 5, located in the Appendix. We experiment with a similar set of baselines as in the case for K, and again observe that the MetaPix model consistently outperforms random initialization or pretraining on all metrics. Also, we see similar generalizability, as the model metatrained for $T = 20$ is able to perform well for other T values at test time too. The ability for MetaPix to generalize across K and T implies cost-effective strategies for training. The computational cost for training a meta-learner is dominated by fine-tuning, which scales linearly with K and T. Training with smaller values of both can result in significant speedups – up to $10\times$ in our experiments.

Variation of meta learning rate ϵ: We also experimented with changing the meta learning rate. At $\epsilon = 0.1$ ($K = 5, T = 200$), we obtained SSIM=0.47, similar to what the pretrained model gets. Using our default $\epsilon = 1.0$, improves performance to 0.51. Hence, a higher meta learning rate was imperative to see improvements with MetaPix.

Only training the generator: We apply Reptile in a GAN setting, where we jointly meta-optimize two networks. We also experimented with freezing one of the networks, specifically the discriminator, to the weights learned during pretraining. For our $K = 5, T = 200, \epsilon = 1.0$ setup, we obtain similar performance as optimizing both, suggesting that a 'universal' discriminator might suffice for meta-learning on GANs.

Temporal coherence in video generation: We generate a video using the pretrained model and MetaPix at $K = 5, T = 200$ setting. We show sample frames in Figure 3 (full video in supplementary). Interestingly, we find that our MetaPix model naturally leads to temporally coherent images, even though it is not explicitly trained for such an objective. This further reinforces our belief that MetaPix learns a much better initialization that is able to quickly adapt to the actor and background appearance from the few samples provided at test time.

4 Conclusion

We have explored the task of quickly and efficiently retargeting human actions from one video to another, given a limited number of samples from the target domain. We formalize this as a few-shot personalization problem, where we first learn a generic Pose2Im generative model on large amounts of data, and then specialize it to a small amount of target frames via finetuning. We propose MetaPix, which repurposes a first-order meta-learning algorithm, Reptile, to adversarially meta-optimize both the generator and discriminator in the Pose2Im network. We experiment with it on in-the-wild YouTube videos, and find that MetaPix significantly outperforms widely-used approaches for pretraining, while generating temporally coherent videos.

References

[1] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. Guttag. Synthesizing images of humans in unseen poses. In *CVPR*, 2018. 1, 2

[2] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros. Everybody dance now. *arXiv preprint arXiv:1808.07371*, 2018. 1, 2, 3

[3] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017. 2

[4] D. P. Kingma and J. L. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 2

[5] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool. Pose guided person image generation. In *NeurIPS*, 2017. 1

[6] A. Nichol, J. Achiam, and J. Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018. 2

[7] A. Siarohin, E. Sangineto, S. Lathuilière, and N. Sebe. Deformable gans for pose-based human image generation. In *CVPR*, 2018. 1

[8] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018. 2, 3, 6

[9] Y. Zhou, Z. Wang, C. Fang, T. Bui, and T. L. Berg. Dance dance generation: Motion transfer for internet videos. *arXiv preprint arXiv:1904.00129*, 2019. 2, 3

Algorithm 1 Meta-learning for video re-targeting.

Initialize θ_D, θ_G from pretrained weights
for iteration = $1, 2, \ldots$ **do**
  Sample K pose image pairs from the same shot randomly
  Compute $\widetilde{\theta_D}, \widetilde{\theta_G} = \text{Pix2PixHD}_K^T(\theta_D, \theta_G)$, for K images and T iterations
  Update $\theta_D = \theta_D - \epsilon(\widetilde{\theta_D} - \theta_D)$
  Update $\theta_G = \theta_G - \epsilon(\widetilde{\theta_G} - \theta_G)$
end for

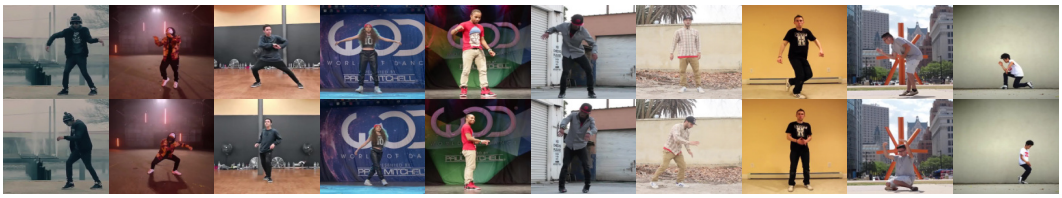

Figure 4: **Training data.** Frames from the additional training data we collected. We download 10 videos from YouTube, distinct from the ones used for personalization and evaluation.

Appendix

Implementation Details. We implement MetaPix by building upon a public Pix2PixHD implementation[1] in PyTorch, and perform all experiments on a 4 TITAN-X or GTX 1080Ti GPU node. We follow the hyperparameter setup as proposed in [8]. We represent the pose using a multi-channel heatmap image, and input and output are 512×512px RGB images. The generator consists of 16 convolutional and deconvolutional layers, and is trained with a equally weighted combination of GAN, Feature Matching, and VGG losses. Initially, we pre-train the model on a large corpus of videos to learn a generic Pose2Im model as described in Section 3. During this pretraining stage, the model is trained on all of the training frames for 10 epochs using learning rate of 0.0002 and batch size of 8 distributed over the 4 GPUs. We experimented with multiple learning rates including $0.2, 0.02, 0.002$; however, we observed that higher learning rates caused the training to diverge. When finetuning for personalization, given K frames and a computational budget T, we train the first $\frac{T}{2}$ iterations using a constant learning rate of 0.0002, and the remaining iterations using a linear decay to 0, following [8]. The batch size is fixed to 8, and for $K < 8$, we repeat the frames to get 8 images for the batch. For the metalearning, we set the meta learning rate, $\epsilon = 1$ with a linear decay to 0, and train 300 meta-iterations. We also experiment with meta learning rate, $\epsilon = 0.1$, however, was much slower to converge. To potentially stabilize metatraining, we experiment with differing numbers of updates to the generator and discriminator during iterations of Alg. 1, as well as simplified objective functions. Recall that the GAN loss adds significant complexity due to the presence of a discriminator that need also be adversarially finetuned. In total, our metalearning takes 1 day of training time on 4 GPUs. We will release the source code for details.

[1] https://github.com/NVIDIA/pix2pixHD/

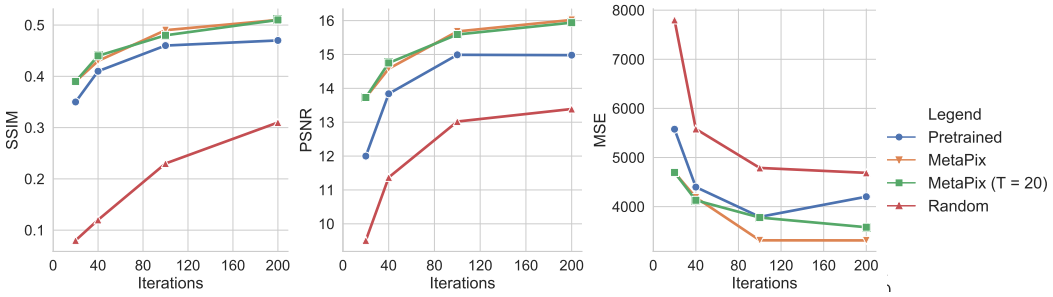

Figure 5: **Personalization after T iterations.** We compare performance on increasing the computational budget for personalization. As expected all initializations improve with T, though MetaPix consistently outperforms random or pretraining. Again we see strong generalizability, as a MetaPix model trained for $T = 20$ performs well at other T values used at test time.

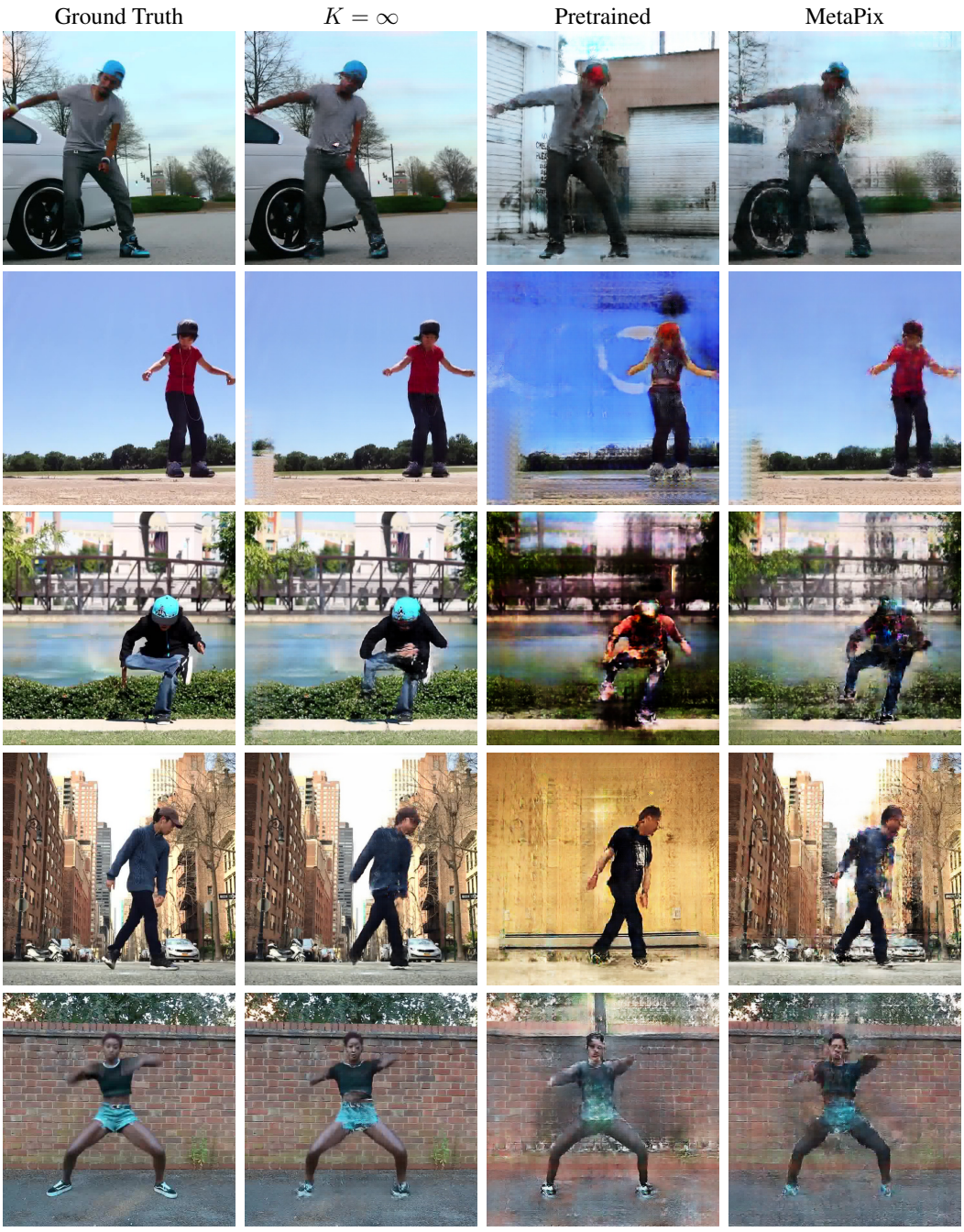

Figure 6: **Qualitative comparison.** Here we compare our MetaPix model with the baselines. The $K = \infty$ model is an upper bound experiment, learning from all the frames available for personalization. The last two columns compare the constrained setting, where at test time only $K = 5$ frames are used for personalization, over $T = 200$ iterations. In the hard cases as shown here, the pretrained model tends to copy over people or background from the training videos when it has not seen the poses in the finetuning set, whereas the MetaPix trained model is much better at learning about the clothing colors, background etc from those few samples.