

Two-Stream Convolutional Networks for Dynamic Texture Synthesis

Matthew Tesfaldet Marcus A. Brubaker

Department of Electrical Engineering and Computer Science
York University, Toronto

{mtesfald, mab}@eecs.yorku.ca

Konstantinos G. Derpanis

Department of Computer Science
Ryerson University, Toronto

kosta@scs.ryerson.ca

Abstract

We introduce a two-stream model for dynamic texture synthesis. Our model is based on pre-trained convolutional networks (ConvNets) that target two independent tasks: (i) object recognition, and (ii) optical flow prediction. Given an input dynamic texture, statistics of filter responses from the object recognition ConvNet encapsulate the per-frame appearance of the input texture, while statistics of filter responses from the optical flow ConvNet model its dynamics. To generate a novel texture, a randomly initialized input sequence is optimized to match the feature statistics from each stream of an example texture. Inspired by recent work on image style transfer and enabled by the two-stream model, we also apply the synthesis approach to combine the texture appearance from one texture with the dynamics of another to generate entirely novel dynamic textures. We show that our approach generates novel, high quality samples that match both the framewise appearance and temporal evolution of input texture. Finally, we quantitatively evaluate our texture synthesis approach with a thorough user study.

1. Introduction

Many common temporal visual patterns are naturally described by the ensemble of appearance and dynamics (*i.e.*, temporal pattern variation) of their constituent elements. Examples of such patterns include fire, fluttering vegetation, and wavy water. Understanding and characterizing these temporal patterns has long been a problem of interest in human perception, computer vision, and computer graphics. These patterns have been previously studied under a variety of names, including turbulent-flow motion [17], temporal textures [30], time-varying textures [3], dynamic textures [8], textured motion [45] and spacetime textures [7]. Here, we adopt the term “dynamic texture”. In this work, we propose a factored analysis of dynamic textures in terms of appearance and temporal dynamics. This factorization is then used to enable dynamic texture synthesis which, based on

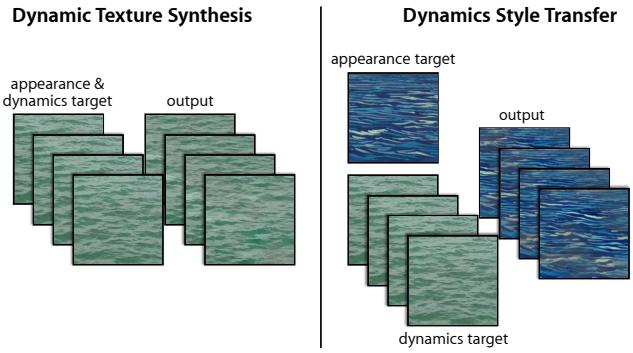


Figure 1: Dynamic texture synthesis. (left) Given an input dynamic texture as the target, our two-stream model is able to synthesize a novel dynamic texture that preserves the target’s appearance and dynamics characteristics. (right) Our two-stream approach enables synthesis that combines the texture appearance from one target with the dynamics from another, resulting in a composition of the two.

example texture inputs, generates a novel dynamic texture instance. It also enables a novel form of style transfer where the target appearance and dynamics can be taken from different sources as shown in Fig. 1.

Our model is constructed from two convolutional networks (ConvNets), an appearance stream and a dynamics stream, which have been pre-trained for object recognition and optical flow prediction, respectively. Similar to previous work on spatial textures [13, 19, 33], we summarize an input dynamic texture in terms of a set of spatiotemporal statistics of filter outputs from each stream. The appearance stream models the per frame appearance of the input texture, while the dynamics stream models its temporal dynamics. The synthesis process consists of optimizing a randomly initialized noise pattern such that its spatiotemporal statistics from each stream match those of the input texture. The architecture is inspired by insights from human perception and neuroscience. In particular, psychophysical studies [6] show that humans are able to perceive the structure of a dynamic texture even in the absence of appearance cues, suggesting that the two streams are effectively inde-

pendent. Similarly, the two-stream hypothesis [16] models the human visual cortex in terms of two pathways, the ventral stream (involved with object recognition) and the dorsal stream (involved with motion processing).

In this paper, our two-stream analysis of dynamic textures is applied to texture synthesis. We consider a range of dynamic textures and show that our approach generates novel, high quality samples that match both the frame-wise appearance and temporal evolution of an input example. Further, the factorization of appearance and dynamics enables a novel form of style transfer, where dynamics of one texture are combined with the appearance of a different one, *cf.* [14]. This can even be done using a single image as an appearance target, which allows static images to be animated. Finally, we validate the perceived realism of our generated textures through an extensive user study.

2. Related work

There are two general approaches that have dominated the texture synthesis literature: non-parametric sampling approaches that synthesize a texture by sampling pixels of a given source texture [10, 26, 37, 47], and statistical parametric models. As our approach is an instance of a parametric model, here we focus on these approaches.

The statistical characterization of visual textures was introduced in the seminal work of Julesz [23]. He conjectured that particular statistics of pixel intensities were sufficient to partition spatial textures into metameristic (*i.e.*, perceptually indistinguishable) classes. Later work leveraged this notion for texture synthesis [19, 33]. In particular, inspired by models of the early stages of visual processing, statistics of (handcrafted) multi-scale oriented filter responses were used to optimize an initial noise pattern to match the filter response statistics of an input texture. More recently, Gatys et al. [13] demonstrated impressive results by replacing the linear filter bank with a ConvNet that, in effect, served as a proxy for the ventral visual processing stream. Textures are modelled in terms of the correlations between filter responses within several layers of the network. In subsequent work, this texture model was used in image style transfer [14], where the style of one image was combined with the image content of another to produce a new image. Ruder et al. [36] extended this model to video by using optical flow to enforce temporal consistency of the resulting imagery.

Variants of linear autoregressive models have been studied [42, 8] that jointly model appearance and dynamics of the spatiotemporal pattern. More recent work has considered ConvNets as a basis for modelling dynamic textures. Xie et al. [48] proposed a spatiotemporal generative model where each dynamic texture is modelled as a random field defined by multiscale, spatiotemporal ConvNet filter responses and dynamic textures are realized by sampling the model. Unlike our current work, which assumes pretrained

fixed networks, this approach requires the ConvNet weights to be trained using the input texture prior to synthesis.

A recent preprint [12] described preliminary results extending the framework of Gatys et al. [13] to model and synthesize dynamic textures by computing a Gram matrix of filter activations over a small temporal window. In contrast, our two stream filtering architecture is more expressive as our dynamics stream is specifically tuned to spatiotemporal dynamics. Moreover, as will be demonstrated, the factorization in terms of appearance and dynamics enables a novel form of style transfer, where the dynamics of one pattern are transferred to the appearance of another to generate an entirely new dynamic texture. To the best of our knowledge, we are the first to demonstrate this form of style transfer.

The recovery of optical flow from temporal imagery has long been studied in computer vision. Traditionally, it has been addressed by handcrafted approaches *e.g.*, [20, 29, 35]. Recently, ConvNet approaches [9, 34, 21, 49] have been demonstrated as viable alternatives. Most closely related to our approach are energy models of visual motion [2, 18, 39, 31, 7, 25] that have been motivated and studied in a variety of contexts, including computer vision, visual neuroscience, and visual psychology. Given an input image sequence, these models consist of an alternating sequence of linear and non-linear operations that yield a distributed representation (*i.e.*, implicitly coded) of pixelwise optical flow. Here, an energy model motivates the representation of observed dynamics which is then encoded as a ConvNet.

3. Technical approach

Our proposed two-stream approach consists of an appearance stream, representing the static (texture) appearance of each frame, and a dynamics stream, representing temporal variations between frames. Each stream consists of a ConvNet whose activation statistics are used to characterize the dynamic texture. Synthesizing a dynamic texture is formulated as an optimization problem with the objective of matching the activation statistics. Our dynamic texture synthesis approach is summarized in Fig. 2 and the individual pieces are described in turn in the following sections.

3.1. Texture model: Appearance stream

The appearance stream follows the spatial texture model introduced by Gatys et al. [13] which we briefly review here. The key idea is that feature correlations in a ConvNet trained for object recognition capture texture appearance. We use the same publicly available normalized VGG-19 network [40] used by Gatys et al. [13].

To capture the appearance of an input dynamic texture, we first perform a forward pass with each frame of the image sequence through the ConvNet and compute the feature activations, $\mathbf{A}^{lt} \in \mathbb{R}^{N_l \times M_l}$, for various levels in the network, where N_l and M_l denote the number of filters and

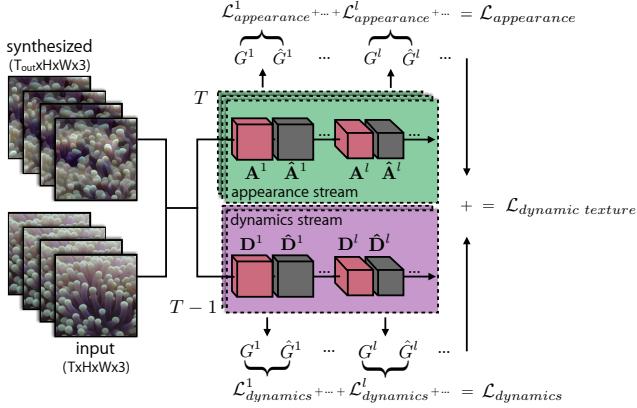


Figure 2: Two-stream dynamic texture generation. Sets of Gram matrices represent a texture’s appearance and dynamics. Matching these statistics allows for the generation of novel textures as well as style transfer between textures.

the number of spatial locations of layer l at time t , respectively. The correlations of the filter responses in a particular layer are averaged over the frames and encapsulated by a Gram matrix, $\mathbf{G}^l \in \mathbb{R}^{N_l \times N_l}$, whose entries are given by $G_{ij}^l = \frac{1}{TN_l M_l} \sum_{t=1}^T \sum_{k=1}^{M_l} A_{ik}^{lt} A_{jk}^{lt}$, where T denotes the number of input frames and A_{ik}^{lt} denotes the activation of feature i at location k in layer l on the target frame t . The synthesized texture appearance is similarly represented by a Gram matrix, $\hat{\mathbf{G}}^{lt} \in \mathbb{R}^{N_l \times N_l}$, whose activations are given by $\hat{G}_{ij}^{lt} = \frac{1}{N_l M_l} \sum_{k=1}^{M_l} \hat{A}_{ik}^{lt} \hat{A}_{jk}^{lt}$, where \hat{A}_{ik}^{lt} denotes the activation of feature i at location k in layer l on the synthesized frame t . The appearance loss, $\mathcal{L}_{\text{appearance}}$, is then defined as the temporal average of the mean squared error between the Gram matrix of the input texture and that of the generated texture computed at each frame:

$$\mathcal{L}_{\text{appearance}} = \frac{1}{L_{\text{app}} T_{\text{out}}} \sum_{t=1}^{T_{\text{out}}} \sum_l \|\mathbf{G}^l - \hat{\mathbf{G}}^{lt}\|_F^2, \quad (1)$$

where L_{app} is the number of layers used to compute Gram matrices, T_{out} is the number of frames being generated in the output, and $\|\cdot\|_F$ is the Frobenius norm. Consistent with previous work [13], we compute Gram matrices on the following layers: *conv1_1*, *pool1*, *pool2*, *pool3*, and *pool4*.

3.2. Texture model: Dynamics stream

There are three primary goals in designing our dynamics stream. First, the activations of the network must represent the temporal variation of the input pattern. Second, the activations should be largely invariant to the appearance of the images which should be characterized by the appearance stream described above. Finally, the representation must be differentiable to enable synthesis. By analogy to the appearance stream, an obvious choice is a ConvNet architec-

ture suited for computing optical flow (*e.g.*, [9, 21]) which is naturally differentiable. However, with most such models it is unclear how invariant their layers are to appearance. Instead, we propose a novel network architecture which is motivated by the spacetime-oriented energy model [7, 39].

In motion energy models, the velocity of image content (*i.e.*, motion) is interpreted as a three-dimensional orientation in the x - y - t spatiotemporal domain [2, 11, 18, 39, 46]. In the frequency domain, the signal energy of a translating pattern can be shown to lie on a plane through the origin where the slant of the plane is defined by the velocity of the pattern. Thus, motion energy models attempt to identify this orientation-plane (and hence the patterns velocity) via a set of image filtering operations. More generally the constituent spacetime orientations for a spectrum of common visual patterns (including translation and dynamic textures) can serve as a basis for describing the temporal variation of an image sequence [7]. This suggests that motion energy models may form an ideal basis for our dynamics stream.

Specifically, we use the spacetime-oriented energy model [7, 39] to motivate our network architecture which we briefly review here; see [7] for a more in-depth description. Given an input video, a bank of oriented 3D filters are applied which are sensitive to a range of spatiotemporal orientations. These filter activations are rectified (squared) and pooled over local regions to make the responses robust to the phase of the input signal, *i.e.*, robust to the alignment of the filter with the underlying image structure. Next, filter activations consistent with the same spacetime orientation are summed. These responses provide a pixelwise distributed measure of which orientations (frequency domain planes) are present in the input. However, these responses are confounded by local image contrast that makes it difficult to determine whether a high response is indicative of the presence of a spacetime orientation or simply due to high image contrast. To address this ambiguity, an L_1 normalization is applied across orientation responses which results in a representation that is robust to local appearance variations but highly selective to spacetime orientation.

Using this model as our basis, we propose the following fully convolutional network [38]. Our ConvNet input is a pair of temporally consecutive greyscale images. Each input pair is first normalized to have zero-mean and unit variance. This step provides a level of invariance to overall brightness and contrast, *i.e.*, global additive and multiplicative signal variations. The first layer consists of 32 3D spacetime convolution filters of size $11 \times 11 \times 2$ (height \times width \times time). Next, a squaring activation function and 5×5 spatial max-pooling (with a stride of one) is applied to make the responses robust to local signal phase. A 1×1 convolution layer follows with 64 filters that combines energy measurements that are consistent with the same orientation. Finally, to remove local contrast dependence, an

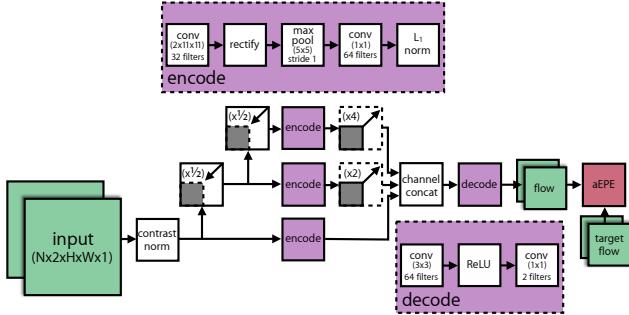


Figure 3: Dynamics stream ConvNet. The ConvNet is based on a spacetime-oriented energy model [7, 39] and is trained for optical flow prediction. Three scales are shown for illustration; in practice five scales were used.

L_1 divisive normalization is applied.

To capture spacetime orientations beyond those capable with the limited receptive fields used in the initial layer, we compute a five-level spatial Gaussian pyramid. Each pyramid level is processed independently with the same spacetime-oriented energy model and then bilinearly upsampled to the original resolution and concatenated.

Prior energy model instantiations (*e.g.*, [2, 7, 39]) used handcrafted filter weights. While a similar approach could be followed here, we opt to learn the weights so that they are better tuned to natural imagery. To train the network weights, we add additional decoding layers that take the concatenated distributed representation and apply a 3×3 convolution (with 64 filters), ReLU activation, and a 1×1 convolution (with 2 filters) that yields a two channel output encoding the optical flow directly. The proposed architecture is illustrated in Fig. 3.

For training, we use the standard average endpoint error (aEPE) flow metric (*i.e.*, L_2 norm) between the predicted flow and the ground truth flow as the loss. Since no large-scale flow dataset exists that captures natural imagery with groundtruth flow, we take an unlabeled video dataset and apply an existing flow estimator [35] to estimate optical flow for training, *cf.* [43]. For training data, we used videos from the UCF101 dataset [41] with geometric and photometric data augmentations similar to those used by FlowNet [9], and optimized the aEPE loss using Adam [24]. Inspection of the learned filters in the initial layer showed evidence of spacetime-oriented filters, consistent with the handcrafted filters used in previous work [7].

Similar to the appearance stream, filter response correlations in a particular layer of the dynamics stream are averaged over the number of image frame pairs and encapsulated by a Gram matrix, $\mathbf{G}^l \in \mathbb{R}^{N_l \times N_l}$, whose entries are given by $G_{ij}^l = \frac{1}{(T-1)N_l M_l} \sum_{t=1}^{T-1} \sum_{k=1}^{M_l} D_{ik}^{lt} D_{jk}^{lt}$, where D_{ik}^{lt} denotes the activation of feature i at location k

in layer l on the target frames t and $t + 1$. The dynamics of the synthesized texture is represented by a Gram matrix of filter response correlations computed separately for each pair of frames, $\hat{\mathbf{G}}^{lt} \in \mathbb{R}^{N_l \times N_l}$, with entries $\hat{G}_{ij}^{lt} = \frac{1}{N_l M_l} \sum_{k=1}^{M_l} \hat{D}_{ik}^{lt} \hat{D}_{jk}^{lt}$, where \hat{D}_{ik}^{lt} denotes the activation of feature i at location k in layer l on the synthesized frames t and $t + 1$. The dynamics loss, $\mathcal{L}_{\text{dynamics}}$, is defined as the average of the mean squared error between the Gram matrices of the input texture and those of the generated texture:

$$\mathcal{L}_{\text{dynamics}} = \frac{1}{L_{\text{dyn}}(T_{\text{out}} - 1)} \sum_{t=1}^{T_{\text{out}}-1} \sum_l \|\mathbf{G}^l - \hat{\mathbf{G}}^{lt}\|_F^2, \quad (2)$$

where L_{dyn} is the number of ConvNet layers being used in the dynamics stream.

Here we propose to use the output of the concatenation layer, where the multiscale distributed representation of orientations is stored, as the layer to compute the Gram matrix. While it is tempting to use the predicted flow output from the network, this generally yields poor results as shown in our evaluation. Due to the complex, temporal variation present in dynamic textures, they contain a variety of local spacetime orientations rather than a single dominant orientation. As a result, the flow estimates will tend to be an average of the underlying orientation measurements and consequently not descriptive. A comparison between the texture synthesis results using the concatenation layer and the predicted flow output is provided in Sec. 4.

3.3. Texture generation

The overall dynamic texture loss consists of the combination of the appearance loss, Eq. (1), and the dynamics loss, Eq. (2):

$$\mathcal{L}_{\text{dynamic texture}} = \alpha \mathcal{L}_{\text{appearance}} + \beta \mathcal{L}_{\text{dynamics}}, \quad (3)$$

where α and β are the weighting factors for the appearance and dynamics content, respectively. Dynamic textures are implicitly defined as the (local) minima of this loss. Textures are generated by optimizing Eq. (3) with respect to the spacetime volume, *i.e.*, the pixels of the video. Variations in the resulting texture are found by initializing the optimization process using IID Gaussian noise. Consistent with previous work [13], we use L-BFGS [28] optimization.

Naive application of the outlined approach will consume increasing amounts of memory as the temporal extent of the dynamic texture grows; this makes it impractical to generate longer sequences. Instead, long sequences can be incrementally generated by separating the sequence into subsequences and optimizing them sequentially. This is realized by initializing the first frame of a subsequence as the last frame from the previous subsequence and keeping it fixed throughout the optimization. The remaining frames

of the subsequence are initialized randomly and optimized as above. This ensures temporal consistency across synthesized subsequences and can be viewed as a form of coordinate descent for the full sequence objective. The flexibility of this framework allows other texture generation problems to be handled simply by altering the initialization of frames and controlling which frames or frame regions are updated.

4. Experimental results

The goal of (dynamic) texture synthesis is to generate samples that are indistinguishable from the real input target texture by a human observer. In this section, we present a variety of synthesis results including a user study to quantitatively evaluate the realism of our results. Given their temporal nature, our results are best viewed as videos. Our two-stream architecture was implemented using TensorFlow [1]. Results were generated using an NVIDIA Titan X (Pascal) GPU and synthesis times ranged between one to three hours to generate 12 frames with an image resolution of 256×256 . For our full synthesis results and source code, please refer to the supplemental material on the project website: ryersonvisionlab.github.io/two-stream-projpage.

4.1. Dynamic texture synthesis

We applied our dynamic texture synthesis process to a wide range of textures which were selected from the Dyn-Tex [32] database and others we collected in the wild. Included in our supplemental material are synthesized results of nearly 60 different textures that encapsulate a range of phenomena, such as flowing water, waves, clouds, fire, rippling flags, waving plants, and schools of fish. Some sample frames are shown in Fig. 4 but we encourage readers to view the videos to fully appreciate the results. In addition, we performed a comparison with [12] and [48]. Generally, we found our results to be qualitatively comparable or better than these methods. See the supplemental for more details on the comparisons with these methods.

We also generated dynamic textures incrementally, as described in Sec. 3.3. The resulting textures were perceptually indistinguishable from those generated with the batch process. Another extension that we explored were textures with no discernible temporal seam between the last and first frames. Played as a loop, these textures appear to be temporally endless. This was achieved by assuming that the first frame follows the final frame and adding an additional loss for the dynamics stream evaluated on that pair of frames.

Example failure modes of our method are presented in Fig. 6. In general, we find that most failures result from inputs that violate the underlying assumption of a dynamic texture, *i.e.*, the appearance and/or dynamics are not spatiotemporally homogeneous. In the case of the *escalator* example, the long edge structures in the ap-

pearance are not spatially homogeneous, and the dynamics vary due to perspective effects that change the motion from downward to outward. The resulting synthesized texture captures an overall downward motion but lacks the perspective effects and is unable to consistently reproduce the long edge structures. This is consistent with previous observations on static texture synthesis [13] and suggests it is a limitation of the appearance stream.

Another example is the *flag* sequence where the rippling dynamics are relatively homogeneous across the pattern but the appearance varies spatially. As expected, the generated texture does not faithfully reproduce the appearance; however, it does exhibit plausible rippling dynamics. In the supplemental material, we include an additional failure case, *cranberries*, which consists of a swirling pattern. Our model faithfully reproduces the appearance but is unable to capture the spatially varying dynamics. Interestingly, it still produces a result which is statistically indistinguishable from real in our user study discussed below.

Appearance vs. dynamics streams We sought to verify that the appearance and dynamics streams were capturing complementary information. To validate that the texture generation of multiple frames would not induce dynamics consistent with the input, we generated frames starting from randomly generated noise but only using the appearance statistics and corresponding loss, *i.e.*, Eq. 1. As expected, this produced frames that were valid textures but with no coherent dynamics present. Results for a sequence containing a school of fish are shown in Fig. 5; to examine the dynamics, see *fish* in the supplemental material.

Similarly, to validate that the dynamics stream did not inadvertently include appearance information, we generated videos using the dynamics loss only, *i.e.*, Eq. 2. The resulting frames had no visible appearance and had an extremely low dynamic range, *i.e.*, the standard deviation of pixel intensities was 10 for values in $[0, 255]$. This indicates a general invariance to appearance and suggests that our two-stream dynamic texture representation has factored appearance and dynamics, as desired.

4.2. User study

Quantitative evaluation for texture synthesis is a particularly challenging task as there is no single correct output when synthesizing new samples of a texture. Like in other image generation tasks (*e.g.*, rendering), human perception is ultimately the most important measure. Thus, we performed a user study to evaluate the perceived realism of our synthesized textures.

Similar to previous image synthesis work (*e.g.*, [5]), we conducted a perceptual experiment with human observers to quantitatively evaluate our synthesis results. We employed a forced-choice evaluation on Amazon Mechanical

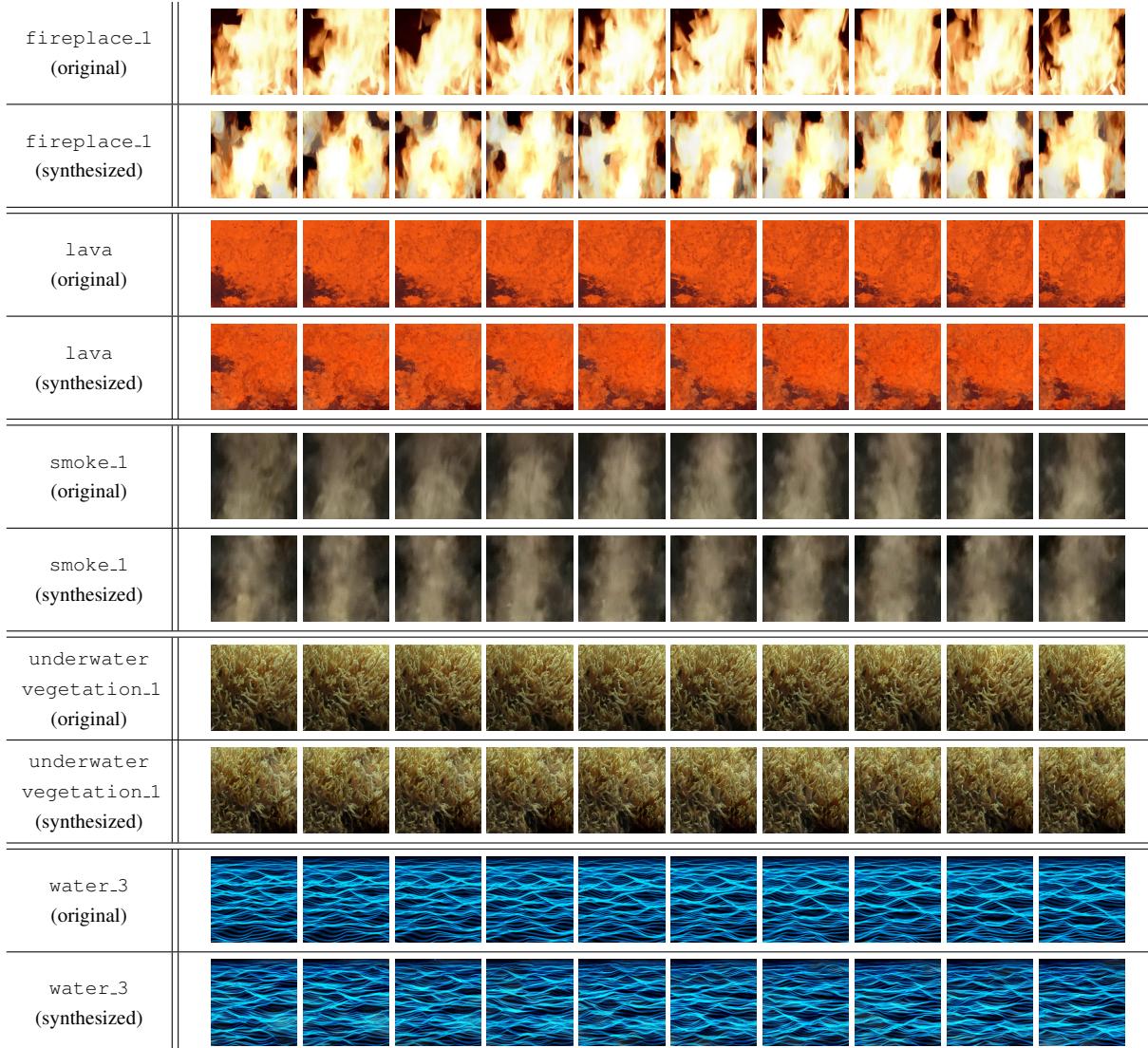


Figure 4: Dynamic texture synthesis success examples. Names correspond to files in the supplemental material.

Turk (AMT) with 200 different users. Each user performed 59 pairwise comparisons between a synthesized dynamic texture and its target. Users were asked to choose which appeared more realistic after viewing the textures for an exposure time sampled randomly from discrete intervals between 0.3 and 4.8 seconds. Measures were taken to control the experimental conditions and minimize the possibility of low quality data. See the supplemental material for further experimental details of our user study.

For comparison, we constructed a baseline by using the flow decode layer in the dynamics loss of Eq. 2. This corresponds with attempting to mimic the optical flow statistics of the texture directly. Textures were synthesized with this model and the user study was repeated with an additional 200 users. To differentiate between the models, we label

“Flow decode layer” and “Concat layer” in the figures to describe our baseline and final model, respectively.

The results of this study are summarized in Fig. 7 which shows user accuracy in differentiating real versus generated textures as a function of time for both methods. Overall, users are able to correctly identify the real texture $66.1\% \pm 2.5\%$ of the time for brief exposures of 0.3 seconds. This rises to $79.6\% \pm 1.1\%$ with exposures of 1.2 seconds and higher. Note that “perfect” synthesis results would have an accuracy of 50%, indicating that users were unable to differentiate between the real and generated textures and higher accuracy indicating less convincing textures.

The results clearly show that the use of the concatenation layer activations is far more effective than the flow decode layer. This is not surprising as optical flow alone is known

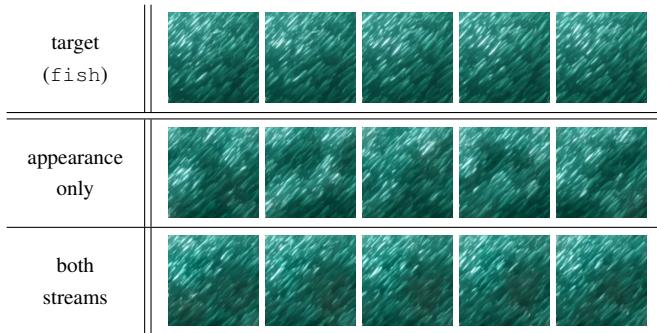


Figure 5: Dynamic texture synthesis versus texture synthesis. (top row) Target texture. (middle) Texture synthesis without dynamics constraints shows consistent per-frame appearance but no temporal coherence. (bottom) Including both streams induces consistent appearance and dynamics.

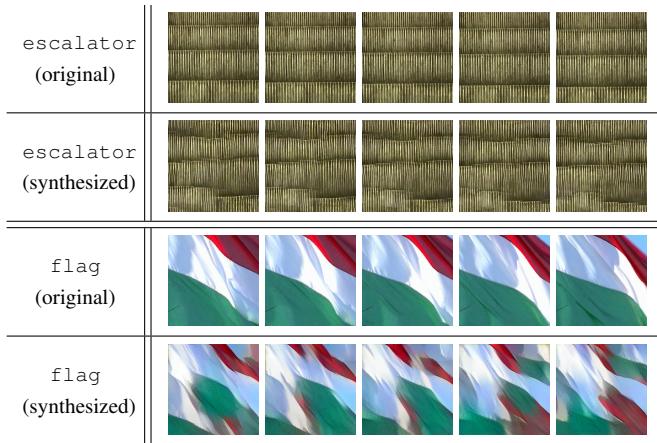


Figure 6: Dynamic texture synthesis failure examples. In these cases, the failures are attributed to either the appearance or the dynamics not being homogeneous.

to be unreliable on many textures, particularly those with transparency or chaotic motion (*e.g.*, water, smoke, flames, etc.). Also evident in these results is the time-dependant nature of perception for textures from both models. Users’ ability to identify the generated texture improved as exposure times increased to 1.2 seconds and remained relatively flat for longer exposures.

To better understand the performance of our approach, we grouped and analyzed the results in terms of appearance and dynamics characteristics. For appearance we used the taxonomy presented in [27] and grouped textures as either regular/near-regular (*e.g.*, periodic tiling and brick wall), irregular (*e.g.*, a field of flowers), or stochastic/near-stochastic (*e.g.*, tv static or water). For dynamics we grouped textures as either spatially-consistent (*e.g.*, closeup of rippling sea water) or spatially-inconsistent (*e.g.*, rippling sea water juxtaposed with translating clouds in the sky). Re-

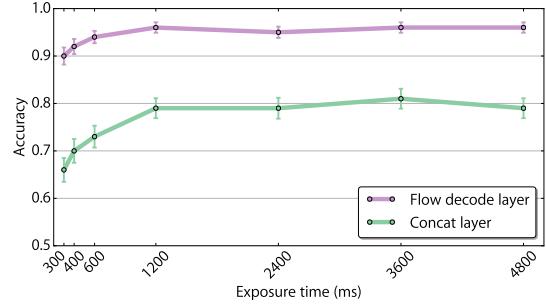


Figure 7: Time-limited pairwise comparisons across all textures with 95% statistical confidence intervals.

sults based on these groupings can be seen in Fig. 8.

A full breakdown of the user study results by texture and grouping can be found in the supplemental material. Here we discuss some of the overall trends. Based on appearance it is clear that textures with large-scale spatial consistencies (regular, near-regular, and irregular textures) tend to perform poorly. Examples being *flag* and *fountain_2* with user accuracies of $98.9\% \pm 1.6\%$ and $90.8\% \pm 4.3\%$ averaged across all exposures, respectively. This is not unexpected and is a fundamental limitation of the local nature of the Gram matrix representation used in the appearance stream which was observed in static texture synthesis [13]. In contrast, stochastic and near-stochastic textures performed significantly better as their smaller-scale local variations are well captured by the appearance stream, for instance *water_1* and *lava* which had average accuracies of $53.8\% \pm 7.4\%$ and $55.6\% \pm 7.4\%$, respectively, making them both statistically indistinguishable from real.

In terms of dynamics, we find that textures with spatially-consistent dynamics (*e.g.*, *tv_static*, *water_**, and *calm_water_**) perform significantly better than those with spatially-inconsistent dynamics (*e.g.*, *candle_flame*, *fountain_2*, and *snake_**), where the dynamics drastically differ across spatial locations. For example, *tv_static* and *calm_water_6* have average accuracies of $48.6\% \pm 7.4\%$ and $63.2\% \pm 7.2\%$, respectively, while *candle_flame* and *snake_5* have average accuracies of $92.4\% \pm 4\%$ and $92.1\% \pm 4\%$, respectively. Overall, our model is capable of reproducing a full spectrum of spatially-consistent dynamics. However, as the appearance shifts from containing small-scale spatial consistencies to containing large-scale consistencies, performance degrades. This was evident in the user study where the best-performing textures typically consisted of a stochastic or near-stochastic appearance with spatially-consistent dynamics. In contrast the worst-performing textures consisted of regular, near-regular, or irregular appearance with spatially-inconsistent dynamics.

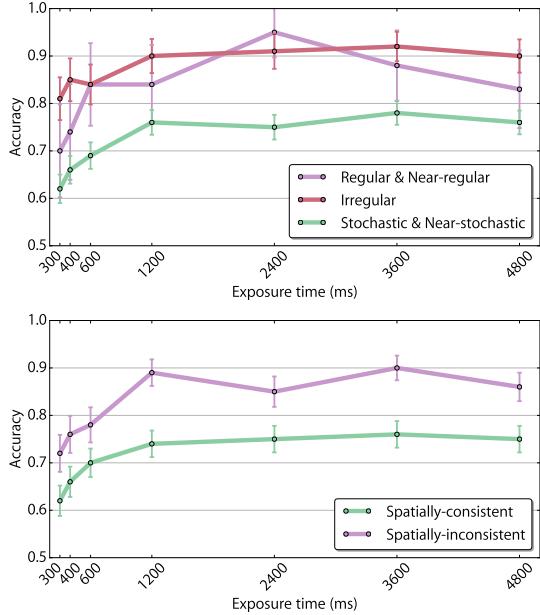


Figure 8: Time-limited pairwise comparisons across all textures, grouped by appearance (top) and dynamics (bottom). Shown with 95% statistical confidence intervals.

4.3. Dynamics style transfer

The underlying assumption of our model is that appearance and dynamics of texture can be factorized. As such, it should allow for the transfer of the dynamics of one texture onto the appearance of another. This has been explored previously for artistic style transfer [4, 15] with static imagery. We accomplish this with our model by performing the same optimization as above, but with the target Gram matrices for appearance and dynamics computed from different textures.

A dynamics style transfer result is shown in Fig. 9 (top), using two real videos. Additional examples are available in the supplemental material. We note that when performing dynamics style transfer it is important that the appearance structure be similar in scale and semantics, otherwise, the generated dynamic textures will look unnatural. For instance, transferring the dynamics of a flame onto a water scene will generally produce implausible results.

We can also apply the dynamics of a texture to a static input image, as the target Gram matrices for the appearance loss can be computed on just a single frame. This allows us to effectively animate regions of a static image. The result of this process can be striking and is visualized in Fig. 9 (bottom), where the appearance is taken from a painting and the dynamics from a real world video.

5. Discussion and summary

In this paper, we presented a novel, two-stream model of dynamic textures using ConvNets to represent the appear-

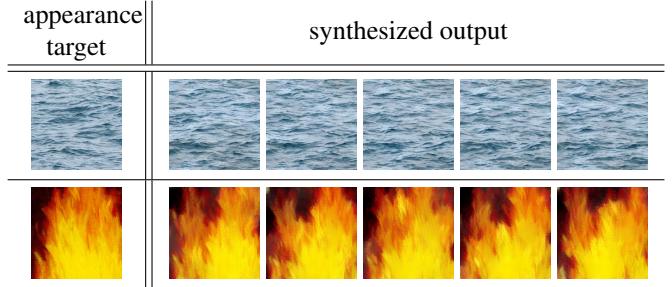


Figure 9: Dynamics style transfer. (top row) Appearance of still water was used with the dynamics of a different water dynamic texture (water_4). (bottom row) The appearance of a painting of fire was used with the dynamics of a real fire (fireplace_1). Animated results and additional examples are available in the supplemental material.

ance and dynamics. We applied this model to a variety of dynamic texture synthesis tasks and showed that, so long as the input textures are generally true dynamic textures, *i.e.*, have spatially invariant statistics and spatiotemporally invariant dynamics, the resulting synthesized textures are compelling. This was validated both qualitatively and quantitatively through a large user study. Further, we showed that the two-stream model enabled dynamics style transfer, where the appearance and dynamics information from different sources can be combined to generate a novel texture.

We have explored this model thoroughly and found a few limitations which we leave as directions for future work. First, much like has been reported in recent image style transfer work [14], we have found that high frequency noise and chromatic aberrations are a problem in generation. Another issue that arises is the model fails to capture textures with spatially-variant appearance, (*e.g.*, flag in Fig. 6) and spatially-inconsistent dynamics (*e.g.*, escalator in Fig. 6). By collapsing the local statistics into a Gram matrix, the spatial and temporal organization is lost. Simple post-processing methods may alleviate some of these issues but we believe that they also point to a need for a better representation. Beyond addressing these limitations, a natural next step would be to extend the idea of a factorized representation into feed-forward generative networks that have found success in static image synthesis, *e.g.*, [22, 44].

Acknowledgements MT is supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) Canadian Graduate Scholarship. KGD and MAB are supported by NSERC Discovery Grants. This research was undertaken as part of the Vision: Science to Applications program, thanks in part to funding from the Canada First Research Excellence Fund.

References

- [1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. [5](#)
- [2] E. H. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion. *JOSA-A*, 2(2):284–299, 1985. [2, 3, 4](#)
- [3] Z. Bar-Joseph, R. El-Yaniv, D. Lischinski, and M. Werman. Texture mixing and texture movie synthesis using statistical learning. *T-VCG*, 7(2):120–135, 2001. [1](#)
- [4] A. J. Champandard. Semantic style transfer and turning two-bit doodles into fine artworks. *arXiv:1603.01768*, 2016. [8](#)
- [5] Q. Chen and V. Koltun. Photographic image synthesis with cascaded refinement networks. In *ICCV*, 2017. [5](#)
- [6] J. E. Cutting. Blowing in the wind: Perceiving structure in trees and bushes. *Cognition*, 12(1):25 – 44, 1982. [1](#)
- [7] K. G. Derpanis and R. P. Wildes. Spacetime texture representation and recognition based on a spatiotemporal orientation analysis. *PAMI*, 34(6):1193–1205, 2012. [1, 2, 3, 4](#)
- [8] G. Doretto, A. Chiuso, Y. N. Wu, and S. Soatto. Dynamic textures. *IJCV*, 51(2):91–109, 2003. [1, 2](#)
- [9] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox. FlowNet: Learning optical flow with convolutional networks. In *ICCV*, pages 2758–2766, 2015. [2, 3, 4](#)
- [10] A. A. Efros and T. K. Leung. Texture synthesis by non-parametric sampling. In *ICCV*, pages 1033–1038, 1999. [2](#)
- [11] M. Fahle and T. Poggio. Visual hyperacuity: Spatiotemporal interpolation in human vision. *Proceedings of the Royal Society of London B: Biological Sciences*, 213(1193):451–477, 1981. [3](#)
- [12] C. M. Funke, L. A. Gatys, A. S. Ecker, and M. Bethge. Synthesising dynamic textures using convolutional neural networks. *arXiv:1702.07006*, 2017. [2, 5](#)
- [13] L. A. Gatys, A. S. Ecker, and M. Bethge. Texture synthesis using convolutional neural networks. In *NIPS*, pages 262–270, 2015. [1, 2, 3, 4, 5, 7](#)
- [14] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *CVPR*, pages 2414–2423, 2016. [2, 8](#)
- [15] L. A. Gatys, A. S. Ecker, M. Bethge, A. Hertzmann, and E. Shechtman. Controlling perceptual factors in neural style transfer. In *CVPR*, 2017. [8](#)
- [16] M. A. Goodale and A. D. Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25, 1992. [2](#)
- [17] D. Heeger and A. Pentland. Seeing structure through chaos. In *IEEE Motion Workshop: Representation and Analysis*, pages 131–136, 1986. [1](#)
- [18] D. J. Heeger. Optical flow using spatiotemporal filters. *IJCV*, 1(4):279–302, 1988. [2, 3](#)
- [19] D. J. Heeger and J. R. Bergen. Pyramid-based texture analysis/synthesis. In *SIGGRAPH*, pages 229–238, 1995. [1, 2](#)
- [20] B. K. P. Horn and B. G. Schunck. Determining optical flow. *A.I.*, 17:185–203, 1981. [2](#)
- [21] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, 2017. [2, 3](#)
- [22] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711, 2016. [8](#)
- [23] B. Julesz. Visual pattern discrimination. *IRE Trans. Information Theory*, 8(2):84–92, 1962. [2](#)
- [24] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980*, 2014. [4](#)
- [25] K. Konda, R. Memisevic, and V. Michalski. Learning to encode motion using spatio-temporal synchrony international conference on learning representation. In *ICLR*, 2014. [2](#)
- [26] V. Kwatra, A. Schödl, I. Essa, G. Turk, and A. Bobick. Graphcut textures: Image and video synthesis using graph cuts. In *SIGGRAPH*, pages 277–286, 2003. [2](#)
- [27] W.-C. Lin, J. Hays, C. Wu, Y. Liu, and V. Kwatra. Quantitative evaluation of near regular texture synthesis algorithms. In *CVPR*, volume 1, pages 427–434, 2006. [7](#)
- [28] D. C. Liu and J. Nocedal. On the limited memory method for large scale optimization. *Mathematical Programming*, 45(3):503–528, 1989. [4](#)
- [29] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, pages 674–679, 1981. [2](#)
- [30] R. Nelson and R. Polana. Qualitative recognition of motion using temporal textures. *CVGIP*, 56(1), 1992. [1](#)
- [31] S. Nishimoto and J. L. Gallant. A three-dimensional spatiotemporal receptive field model explains responses of area mt neurons to naturalistic movies. *Journal of Neuroscience*, 31(41):14551–14564, 2011. [2](#)
- [32] R. Péteri, S. Fazekas, and M. J. Huiskes. DynTex: A Comprehensive Database of Dynamic Textures. *PRL*, 31(12), 2010. [5](#)
- [33] J. Portilla and E. P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *IJCV*, 40(1):49–70, 2000. [1, 2](#)
- [34] A. Ranjan and M. J. Black. Optical Flow Estimation using a Spatial Pyramid Network. In *CVPR*, 2017. [2](#)
- [35] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid. EpicFlow: Edge-preserving interpolation of correspondences for optical flow. In *CVPR*, pages 1164–1172, 2015. [2, 4](#)
- [36] M. Ruder, A. Dosovitskiy, and T. Brox. Artistic style transfer for videos. In *GCPR*, pages 26–36, 2016. [2](#)
- [37] A. Schödl, R. Szeliski, D. Salesin, and I. A. Essa. Video textures. In *SIGGRAPH*, pages 489–498, 2000. [2](#)
- [38] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *PAMI*, 39(4):640–651, 2017. [3](#)

- [39] E. P. Simoncelli and D. J. Heeger. A model of neuronal responses in visual area MT. *Vision Research*, 38(5):743 – 761, 1998. [2](#), [3](#), [4](#)
- [40] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014. [2](#)
- [41] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv:1212.0402*, 2012. [4](#)
- [42] M. Sumer and R. W. Picard. Temporal texture modeling. In *ICIP*, pages 823–826, 1996. [2](#)
- [43] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Deep end2end voxel2voxel prediction. In *CVPR Workshops*, pages 402–409, 2016. [4](#)
- [44] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. S. Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. In *ICML*, pages 1349–1357, 2016. [8](#)
- [45] Y. Wang and S. C. Zhu. Modeling textured motion: Particle, wave and sketch. In *ICCV*, pages 213–220, 2003. [1](#)
- [46] A. B. Watson and A. J. Ahumada. A look at motion in the frequency domain. In *Motion workshop: Perception and representation*, pages 1–10, 1983. [3](#)
- [47] L. Wei and M. Levoy. Fast texture synthesis using tree-structured vector quantization. In *SIGGRAPH*, pages 479–488, 2000. [2](#)
- [48] J. Xie, S.-C. Zhu, and Y. N. Wu. Synthesizing dynamic patterns by spatial-temporal generative convnet. In *CVPR*, 2017. [2](#), [5](#)
- [49] J. J. Yu, A. W. Harley, and K. G. Derpanis. Back to Basics: Unsupervised Learning of Optical Flow via Brightness Constancy and Motion Smoothness. In *ECCVW*, 2016. [2](#)