Customizing First Person Image Through Desired Actions

Shan Su  
University of Pennsylvania  
sushan@seas.upenn.edu

Jianbo Shi  
University of Pennsylvania  
jshi@seas.upenn.edu

Hyun Soo Park  
University of Minnesota  
hspark@umn.edu

Abstract

This paper studies a problem of inverse visual path planning: creating a visual scene from a first person action. Our conjecture is that the spatial arrangement of a first person visual scene is deployed to afford an action, and therefore, the action can be inversely used to synthesize a new scene such that the action is feasible. As a proof-of-concept, we focus on linking visual experiences induced by walking.

A key innovation of this paper is a concept of ActionTunnel—a 3D virtual tunnel along the future trajectory encoding what the wearer will visually experience as moving into the scene. This connects two distinctive first person images through similar walking paths. Our method takes a first person image with a user defined future trajectory and outputs a new image that can afford the future motion. The image is created by combining present and future ActionTunnels in 3D where the missing pixels in adjoining area are computed by a generative adversarial network. Our work can provide a travel across different first person experiences in diverse real world scenes.

1. Introduction

Imagine you have a picture of the scene in front of you, as shown in Scene 1 of Figure 1. You see the road, the parked cars (right), and fence (left). This scene makes you move straight. What if you want to make a right turn half way down the road. How do you imagine the scene so that you can modify the action? We would need to create an intersection for you to turn, and a side street to continue onto. In another example, if we want to enter a building to the left, we would need to create an entrance on the building.

In this paper, we propose a problem of inverse visual path planning, i.e., creating a visual scene from a future path. Our key insight is based on Gibson’s ecological perception: a strong duality between visual scene and action. The spatial arrangement of first person scenes affords actions, and therefore, it is possible to visualize what we would see based on how we will act. A key challenge is to embed visual semantics of a new future action into the present scene, i.e., synthesizing pixels such that the present image can afford future as shown in Composite image of Figure 1.

We address this challenge by leveraging a novel concept of ActionTunnel—a 3D virtual tunnel along the future trajectory encoding what the camera wearer will visually experience as moving into the scene akin to the spacetime tunnel in physics. ActionTunnel connects two distinctive scenes in a spatially consistent way, which allows creating a new image that affords the desired future action. As a proof-of-concept, we focus on linking such visual sensations through walking.

Our method takes a first person RGBD image with a user defined future trajectory and outputs a new image that the desired motion is feasible. We represent an image using ActionTunnel along the future trajectory where its cross section is determined by the walkable pixels on the ground plane. The future visual scene is selected according to the similarity of future trajectories and their surrounding visual context. We combine present and future ActionTunnels in 3D to generate a novel image where the missing pixels in adjoining area are computed by a generative adversarial network. The resulting ActionTunnel is projected onto the camera pose, which produces a continuous transition.
We cast this problem as learning visual context of walking path memory. A visual scene can be decomposed according to short and long term desired paths where we find their transition and geometric alignment of the present and future scenes, and glue them back into a composite image. Why useful? An ability to connect visual sensations is a key design factor of AI because it can index visual scenes in a spatially persistent way. For instance, our visual sensation alignment can generate the contents for virtual reality (VR), providing new supervisionary signals for learning visual semantics across time, and share virtual experiences via first person videos. Also this can close the loop of the planning by perception paradigm, i.e., our work provides perception through planning, which can feedback to the planning. Beyond AI and VR, this work brings a new opportunity to study the human visual memory system, establishing a tight integration of memory into visual perception and motor skill [55].

Contribution To our best knowledge, this is the first paper that addresses connecting first person visual sensations through actions. The core technical contributions include: 1) ActionTunnel representation: we develop a novel visual representation persistent to the 3D spatial structure given future trajectory. 2) Visual sensation alignment and retrieval: our algorithm predicts a scene that produces a plausible transition by aligning ActionTunnel. 3) Gap filling: we use a generative adversarial network to fill the pixels in adjoining area, which connects the ActionTunnel. We demonstrate that our work can provide travel across different first person experiences in diverse real world scenes.

2. Related Work

Connecting human visual sensations through memory system, i.e., how we index and store what we have seen, has been the central theme of neuroscience [53, 54] and cognitive vision science [4, 31, 39]. This has motivated computer vision and graphics research to build a computational visual memory system that connects the humongous image and video data in a semantically meaningful way.

A core machinery behind linking the visual sensations is the space and time registration of visual data, discovering invariant visual patterns that can relate distinctive instances/experiences [9]. Due to its highly structured shape, human faces are one of the earliest targets in visual registrations in computer vision (e.g., eigenface [52]). Such face alignment has been the backbone of reconstructing occluded and noisy face [56], tracking [42] and re-targeting facial expressions [16], and building a progression of a photo album via spatiotemporal registration [17]. Beyond a single object such as a face, aligning spatial context together can strengthen the links between visual sensations [29, 32]. Such contextual visual information is deeply embedded in our daily lives, which provides a strong cue to localize the

image in the world [6, 44], recognize human sketches [23], explore multiple images together [59], and arrange images based on space and time [21]. Notably, 3D spatial registration using structure from motion links a set of unorganized images from different perspectives [46, 47] or videos [2, 50]. Also precise temporal ordering based on visual semantics produces consistent visual experiences. Moving objects in multiple images provides a geometric constraint across time [5], finding a smooth transition between images based on geometry and appearance generates a time-lapse video [30], and a prior knowledge of urban scene structure such as buildings relates historical pictures [43].

Unfortunately, such visual registrations are still not sufficient to describe the way we connect our visual episodic memory. It mainly focuses on knowing a factual event, e.g., where it was, how it looked, and what happened, while our visual memory system delves for remembering the personal experiences through interactions, e.g., what I did and how I felt [51]. We attempt to fill this gap by exploiting first person videos that capture the visual sensation induced by walking actions.

A key property of a first person video is that it can closely capture what is important to the camera wearer through interactions [3, 22], which is highly relevant to episodic visual memories. Two visual signals delineate the first person interactions. First, the actions of the camera wearers link past experiences. The gaze direction tells us about the visual attention while interacting with objects [7, 25] and people [8, 35, 36], and hand manipulation provides a strong cue to recognize activities [7, 26, 28, 40]. Therefore, it is possible to infer what they did, and further identify how much they paid attention [49], which physical force they experienced [34], and how they felt through physiological state estimation [13]. Second, the surrounding visual scenes relates visual sensations similar to Déjà vu experiences. First person visual data are highly structured due to camera placement, anthropometric constraints, and proxemics [11], which allows characterizing visual spatial layout. For instance, unique dynamics of sport scenes can be recognized [18], the second person’s activities with respect to the first person can be inferred [41], and semantic segmentation for body parts can be used to understand the state of the interacting objects [24, 25]. Using such first person characteristics, the future actions can be forecasted via behavioral cloning [33, 45] and visual features can be learned without supervision [1, 15]. These two signals from first person videos are integrated to summarize lifelogging videos [27, 58], identifying an memorable moment [57], editing first person videos [2, 19].

The innovation of our work is to use first person actions as a cue to link between visual sensations without supervision by aligning them through ActionTunnel. This produces a continuous transition from one scene to the other.
3. ActionTunnel Model

We construct an ActionTunnel model—a 3D virtual tunnel along the future trajectory encoding what the camera wearer will visually experience as moving into the scene (Figure 2(a)). This model allows transitioning from one visual scene to the other according to the action (Section 4). The desired property of the ActionTunnel is the ability to edit the present image to embed the visual semantics of the desired future action without manual supervision. We represent ActionTunnel in the 3D proxemic space [11] encoding visual and spatial semantics that affords the walking actions.

3.1. 3D Proxemic Space

We define a 3D proxemic space using a log-cylindrical coordinate that highly emphasizes on the short range area. A 3D point $x = G_o + Xg_x + Yg_y + Zg_z$ in the world coordinate is mapped to $(r, \theta, h)$ such that $r = \log \rho$, $\theta = \text{atan2}(Y, X)$, and $h = Y/r$ where $\rho = \sqrt{X^2 + Y^2}$, $g_x, g_y \in S^2$ are two unit vectors spanning the ground plane, and $g_z = g_x \times g_y$ is its surface normal as shown in Figure 2(a). We align $g_y$ with the instantaneous velocity, $v = C_{t+1} - C_t$, $C \in \mathbb{R}^3$ is the optical center of the first person camera, and $G_o = C - Hg_z$ is the origin of the ground plane where $H$ is the height of the camera wearer.

This proxemic space preserves 3D distance near the camera wearer. The EgoRetinal map [33] (Figure 3)\(^2\) is a special instantiation of this log-cylindrical space, which is its ground plane projection. This reflects both first person coordinate and 3D world coordinate: it does not introduce significant pixel perspective distortion while preserving 3D visual sensation similar to EgoRetinal map.

Note that the camera orientation is not necessarily aligned with the moving direction. In order to learn a first person visual representation invariant to head movement and camera placement, we rectify the camera orientation $R \in SO(3)$ with respect to the ground plane such that $g_z = -r_y$ where $r_y$ is the y-axis of the camera, i.e., the rectified camera orientation is $R = [g_x, -g_y, g_z]^T$.

We warp the image $I$ using a homography induced by $R$, $\bar{I} = I(KRR^T u)$, where $K$ is the intrinsic parameter and $u$ is image coordinate: the camera projection matrix is $P = KR \begin{bmatrix} I & -C \end{bmatrix} \in \mathbb{R}^{3 \times 4}$. This rectified image aligns with the ground plane and instantaneous velocity, which allows us to stabilize jittery first person videos due to severe head movement [33, 48] as shown in Figure 2(b). The all vanishing lines of first person images are perfectly aligned, enabling visual learning efficient where the depth and shape of an object is highly predictable by the pixel location as discussed in Section 4.2.

3.2. ActionTunnel

We define ActionTunnel such that it exhibits the desired property, i.e., embedding future into the present image, through three key ingredients inherited in first person images.

First, ActionTunnel encodes the action associated with a first person image. It is formed around the future trajectory that the person would walk over. We represent the future trajectory in the 3D proxemic space as $T = \begin{bmatrix} t_1^T & \ldots & t_p^T \end{bmatrix}^T$, where $t_i = \begin{bmatrix} r_i & \theta_i & 0 \end{bmatrix}^T$ is the 3D coordinates of person’s foot location at the $i^{th}$ time instant later in the future (Figure 3(b)).

Second, the cross section of ActionTunnel encodes the space that affords the action, i.e., sidewalk. We define the cross section as a lateral spatial extent of a walking trajectory. The lateral cross section, $V =$

---

\(^2\)The EgoRetinal map is a visual representation experienced from first person view but visualized in an overhead bird-eye map, akin to an illustrated tourist map.
Walkable pixels

$$\text{Width of cross section}$$

$$1\text{v}$$

$$2\text{v}$$

Figure 3. We construct ActionTunnel in the 3D proxemic space (log-cylindrical coordinate). Its cross section is computed by the height map of EgoRetinal representation [33].

$$\begin{bmatrix} b_1^T & b_2^T & c_1^T & c_2^T \end{bmatrix}^T$$

is defined by two bottom corners $$b_1, b_2$$ and two ceiling corners $$c_1, c_2$$ (Figure 2(a)). For the bottom corners, we seek the farthest point from the trajectory lower than $$h_{\text{max}}$$ along $$v^\perp$$ where $$v^\perp$$ is the direction perpendicular to the instantaneous velocity $$v_i = t_{i+1} - t_i$$ as shown in Figure 3(a). Formally, $$b_j = \lambda_{\text{max}} v_j^\perp + t_i$$ for $$j = 1, 2$$ where

$$\lambda_{\text{max}} = \arg\max \lambda \left\{ \lambda | \phi(\lambda v_j^\perp + t_i) < h_{\text{max}}, \lambda > 0 \right\}.$$ $$h_{\text{max}}$$ defines the walkable height in the log-cylindrical space, e.g., vehicles or buildings are not walkable. $$v^\perp$$ is a vector on the ground plane perpendicular to the $$v_i$$, i.e., $$v_i^T v^\perp = 0$$, and $$v_i^\perp = -v_2^\perp$$. $$\phi(r, \theta, 0)$$ is the height map from EgoRetinal map [33] as shown in Figure 3(a). The ceiling point is $$\alpha H$$ higher than $$b$$ in the 3D world along $$g_z$$ direction, which is translated to $$c_j = b_j + \left[ 0 \ 0 \ \alpha H/Z_i \right]^T$$ in the log-cylindrical coordinate, where $$\alpha$$ defines the fixed height of the ActionTunnel.

Third, we map the texture from the first person image to the ActionTunnel surfaces defined by adjacent cross sections, $$\mathcal{J}_i$$. We convert the vertices of the ActionTunnel from the proxemic space to the 3D world coordinate, i.e., $X = \exp(r) \cos \theta, Y = \exp(r) \sin \theta, Z = h \exp(r)$, and then, project onto the rectified image plane, i.e., $\lambda u = K \bar{R} (X - C)$ where $X = \begin{bmatrix} X & Y & Z \end{bmatrix}^T$.

In total, the ActionTunnel is composed of the series of cross section vertices and associated texture along the future trajectory $V = \{ V_i, \mathcal{J}_i \}_{i=1}^\infty$. To visualize an image from a novel view, the texture of the ActionTunnel in the 3D proxemic space is projected to the rectified camera pose: $\bar{I} = f_{\text{PROJ}}(\bar{V}; \bar{P})$ where $\bar{P} = K \bar{R} \begin{bmatrix} I & -C \end{bmatrix}$ is the rectified camera projection matrix, or inversely, $V = f_{\text{PROJ}}^{-1}(\bar{I}, \bar{P})$ is projecting from the first person image to the ActionTunnel (Figure 4(b)).

4. Connecting Visual Sensations

As parametrized by an action, the ActionTunnel is a medium that connects two distinctive visual sensations through walking future trajectories. In this section, we recognize a plausible transition between two visual scenes and construct a novel ActionTunnel by combining them, which allows us to create a composite picture using a generative adversarial network. We defer the visual scene retrieval to Section 4.3.

4.1. ActionTunnel Composition

Given ActionTunnel $\mathcal{V}_1 = f_{\text{PROJ}}^{-1}(\mathcal{T}_1; \bar{P}_1)$ to $\mathcal{V}_2 = f_{\text{PROJ}}^{-1}(\mathcal{T}_2; \bar{P}_2)$, we compose a novel transitional image, $\mathcal{I}_c$:

$$\mathcal{I}_c = f_{\text{PROJ}}(\mathcal{V}_c = \mathcal{V}_1 \cup \mathcal{V}_2; \bar{P}_1),$$ (1)

where $\mathcal{V}_c$ is the new ActionTunnel constructed by combining $\mathcal{V}_1$ and $\mathcal{V}_2$. The segmented ActionTunnels are $\mathcal{V}_1 = \{ \mathcal{V}_i \in \mathcal{V}_1 | t_i^r < R - \Delta R \}$ and $\mathcal{V}_2 = \{ \mathcal{V}_i \in \mathcal{V}_2 | t_i^r > R + \Delta R \}$ as shown in Figure 4(a) where a transitional point $R$ with transitional length, $\Delta R$, and $t^r$ is the range coordinate of trajectory point $t$. Note that $\mathcal{V}_1$ to $\mathcal{V}_2$ are aligned such that $\mathcal{V}_1 \parallel \mathcal{V}_2 \perp$ where $\mathcal{V}_2 \perp = (\mathcal{V}_2 \setminus \mathcal{V}_1)$ is the complementary set of $\mathcal{V}_2$ described in Section 4.3. Figure 4(c) illustrates the projection of the ActionTunnel onto the first person camera pose where the 3D scene is pre-aligned, e.g., vanishing lines and movement direction invariant to 3D head pitch angle.

$\mathcal{I}_c$ encodes the key desired property of ActionTunnel: indicating which pixel to change and stay in a spatially persistent way. The pixels in the adjoining area (masked area) are changeable while others are less changeable.

Given $\mathcal{I}_c$, we predict the complete image, $\mathcal{I}_g$ including the texture of the adjoining region using a generative model $G$:

$$\mathcal{I}_g = G(\mathcal{I}_c, z; w_G)$$ (2)

where $G$ predicts a complete image given incomplete image $\mathcal{I}_c$ parametrized by $w_G$ as shown in Figure 4(c). This generative model is composed of encoder $E$ and decoder $E^{-1}$, i.e., $G = E^{-1}(E(\mathcal{I}_c), z)$. $E$ generates a compact visual representation invariant to the location of the missing pixels [37] and $E^{-1}$ reconstructs the complete image with $z \sim \mathcal{N}(0; I)$ which is a random vector drawn from a zero-mean Gaussian distribution. We rotate back the complete image rectified by $\bar{R}$ to the original first person camera pose by $\bar{I}_g = \mathcal{I}_g (K \bar{R} \bar{I}^T K^T u)$.

4.2. Generative Image Predictor

We train the generative model $G$ jointly with a discriminative model $D$ using a generative adversarial network (GAN) [10, 37]. $D$ differentiates a real image from an image generated by $G$, while $G$ is trained to confuse $D$. Such adversarial setting reinforces each other, exhibiting strong predictive power. The following optimization trains these
We generate the training data using Action-visual structure. This allows the network to capture the details of which is beneficial for efficient training and information sharing. Instead, we use the same image to represent two viewpoints, i.e., 

\[ \text{REC} = \min_{w_c} \max_{w_d} \mathcal{L}_{\text{GAN}} + \lambda \mathcal{L}_{\text{REC}}, \]

where \( \mathcal{L}_{\text{GAN}} \) is the generative adversarial loss that measures the confusion of the discriminative network between the real and generated images:

\[ \mathcal{L}_{\text{GAN}} = \mathbb{E}_I [\log(D(I))] + \mathbb{E}_{z \sim \mathcal{N}} [\log(1 - D(G(I; z)))] \]

\( \mathcal{L}_{\text{REC}} \) is the reconstruction error that penalizes the difference of generated image from the corresponding image:

\[ \mathcal{L}_{\text{REC}} = ||M \odot (I - G(I; z))||_1 \]

where \( M \) is a binary mask indicating the valid pixel locations in the rectified image \( I \), induced by a homography \( KR^T \cdot V^T \).

Our GAN is inspired by the design of a deep convolutional generative adversarial network [37, 38], composed by 5 convolutional layers followed by batch normalization and Leaky ReLU as shown in Figure 5. The encoder takes images of size \( 256 \times 256 \times 3 \) and outputs 100 dimensional visual feature \( E(I_c) \). We augment the bypassing connections in the corresponding layers of the encoder/decoder, which is beneficial for efficient training and information sharing. This allows the network to capture the details of visual structure.

**Training data** We generate the training data using ActionTunnel. Note that no ground truth data is available in theory. Instead, we use the same image to represent two visual scenes in Equation (1). Given a rectified image \( \mathcal{T}_s \), we build a masked image, \( \mathcal{T}_m \), by constructing the ActionTunnel and leaving out vertices, i.e., \( \mathcal{T}_m = f_{\text{PROJ}}(\mathcal{T}_m) \) where \( \mathcal{V}_m = \{ V_i \in f_{\text{PROJ}}(\mathcal{T}_s) | \| V_i \|_1 < \| R - \Delta R, R + \Delta R \| \} \). The training data has diverse shapes, sizes, and locations.

As the training and testing distributions are different, i.e., testing from two images while training from the same image, it is possible that predicting \( \mathcal{T}_g \) is challenging. However, precise alignment and retrieval in Section 4.3 allows us to align 3D structure resulting in plausible continuous transitions in practice.

### 4.3. Retrieval

Given a first person image, we find an image taken in different time and space that can produce plausible transition. Two conditions have to be made: 1) future trajectories are similar in particular, near the camera wearer, which provides an immediate continuous transition while long term future can be discounted; 2) 3D spatial layouts such as road, building, and car are aligned rather than appearance itself because motion perception is more dominant than semantic scene recognition where the camera wearer undergoes strong egomotion, e.g., we often perceive “something” on the left without knowing what it is when turning right: “knowing” vs. “remembering” [51].

We define the distance metric, \( D(\cdot, \cdot) \), as follow:

\[ D(I_1, I_2) = \lambda_V D_V(\mathcal{T}_1, \mathcal{T}_2) + \lambda_S D_S(\phi_1, \phi_2) + \lambda_M D_M(V_1, V_2) \]

where \( D_V, D_S, \) and \( D_M \) are visual, spatial, and motion distance, respectively, and \( \lambda_V, \lambda_S, \) and \( \lambda_M \) are their weights.

The visual distance measures the scene difference invariant to translation: \( D_V(\mathcal{T}_1, \mathcal{T}_2) = \| F(\mathcal{T}_1) - F(\mathcal{T}_2) \|^2 \) where \( F \) is a compact visual feature for the rectified image. Our choice of \( F \) is a convolutional neural network [20]. These visual features are highly indicative of visual scene context robust to translation.

The spatial distance measures the 3D scene similarity which requires a precise alignment: \( D_S(\phi_1, \phi_2) = \max_{\Delta \theta} (1 - NCC(\phi_1(r, \Delta \theta), \phi_2(r, 0))) \) where \( NCC(\cdot, \cdot) \) measures the normalized cross-correlation between the height map \( \phi \) from EgoRetinal map. \( D_S \) finds the maximum response across angle, which produces the best alignment.

Lastly, we measure motion difference via ActionTunnel:

\[ D_M = \min_{\Delta X} \sum_{i=1}^F \| V_i^1 + \Delta X - V_i^2 \|^2 \]

where \( V_i^1 \) and \( V_i^2 \) are the vertices of cross sections of first and second ActionTunnels, respectively, and \( F \) is the number of frames in the
trajectory of the ActionTunnel. This also align motion in
terms of translation, $x$.

Using this distance metric, we retrieve and alignment
together. In practice, we align the angle using $D_3$ and then,
the translation using $D_M$. This distance measure encodes
two similarities: trajectory and spatial layout.

Given $k$ nearest neighbors based on $D(I_1, I_2)$, we verify the
plausibility of the scene transition using the discriminative
model, $D$, learned in Section 4.2. $D(I_c)$ computes the
likelihood of being real images, which allows us to rank the
neighbors.

5. Result

We use an egomotion public dataset [33] to evaluate our
work where the 3D future trajectory per image was recon-
structured using their structure from motion [12]. This dataset
includes diverse walking sequences in different locations and
time (13 outdoor such as Park, Campus, and Downtown, 13 indoor scenes such as Mall, IKEA, Costco). The
dataset provides the full physical scale of reconstruction at
5 fps and the disparity map from 100 mm baseline stereo
system. ActionTunnel based on our 3D proxemic space in
Section 3.1 is computed based on their EgoRetinal repre-
sentation. Note that for all evaluations, the training set is
completely isolated from the testing set.

5.1. Quantitative Evaluation

As discussed in Section 4.2, there exists no ground truth
data because a visual scene transition is fundamentally im-
possible in reality. Instead, we present a unique measure
that can evaluate the validity of our method: given an im-
age, $\mathbb{I}$ with missing pixels, $M_f \otimes \mathbb{I}$ where the mask $M_f$
indicating valid pixels, we predict them using the remain-
ing pixels $((1 - M_f) \otimes \mathbb{I})$ and a consecutive image $\mathbb{I}_{\Delta t}$.
We measure reconstruction validity using normalized cross
-correlation (NCC), i.e., $\eta = NCC(M_f \otimes \mathbb{I}, M_f \otimes \mathbb{I}_{\Delta t})$ where $\mathbb{I}_{\Delta t}$ is the reconstructed image. This evaluation metric
differs from the inpainting measure [37] because we allow
additional visual cues from consecutive images. This metric
can directly measure the efficacy of ActionTunnel that auto-
matically recognizes which pixel to change and stay (pixel persisten-
tce).

We compare our method with 5 baseline algorithms. 2D
pasting: We predict the missing pixels by copying the pix-
els from the consecutive frame, i.e., $(1 - M_f) \otimes \mathbb{I} =
(1 - M_f) \otimes \mathbb{I}_{\Delta t}$. If the camera wearer does not move,
this measure will be maximized. 2D GAN filling: The
prediction by the inpainting approach [37] that uses a gen-
erative adversarial network is compared. This work has
been evaluated on Street view and Imagenet, which could
be challenging to predict a first person image due to the
larger variation of visual data caused by head movement.
3D space pasting: We evaluate the effectiveness of know-
ing 3D based on our rectification using the ground plane dis-
cussed in Section 3.1. To eliminate the effect of the walk-
ing action, we do not take into account the instantaneous
velocity, $v$, to define $R$. We predict pixels in the recti-
ified image by copying pixels from the consecutive image,
$(1 - M_f) \otimes \mathbb{I} = (1 - M_f) \otimes \mathbb{I}_{\Delta t}$. Note that the mask is
also rectified, $\mathbb{M}_f$. 3D GAN space filling: Similar to 2D
GAN filling, we perform this in the rectified image where
the vanishing lines of all first person images are aligned.
3D Box with GAN: We use a 3D box representation of the
world [14] to recognize the pixel persistence. This is similar
to our method except for the fact that it does not take into
account actions. For each evaluation, we adjust the shape of
the mask to be consistent across baseline algorithms.

We evaluate our work in terms of two aspects: general-
ization power and missing data.

Generalization power We predict the missing pixels (more
than 60% in image) by changing $\Delta t$, the 3D distance be-
tween two images. This can evaluate how far the algorithm
can benefit from the consecutive image. As increasing $\Delta t$,
the visual similarity is less likely consistent. Also the head
orientation plays a significant role in prediction where 3D
based methods often outperform 2D based methods (2D
pasting, 2D GAN filling). In particular, 2D GAN filling
shows limited performance on first person image prediction
due to unstable head movement. Interestingly, 2D pasting
sometimes outperform 2D GAN, which indicates exploit-
Figure 6. We compare our method with 5 baseline algorithms for generalization power: 2D pasting, 2D GAN filling [37], 3D space pasting, 3D space GAN filling, and 3D box with GAN [14]. We reconstruct a masked image using a consecutive image with ∆t frame difference. Our method consistently outperforms others where stronger prediction is achieved for outdoor scenes which aligns with the prior observations [33].

5.2. Qualitative Evaluation

We apply our method to diverse real world scenes both outdoor and indoor. Figure 8 illustrates a sequence of continuous transitional images from present visual scene to the future as the camera wearer walks over time. To generate the transition, we set the transitional point in 3D to compute R per time instant, which allows us to travel one space to the other that affords the different future trajectory. The color of trajectory indicates the transitional point where ActionTunnels are combined. Our method produces a creative transition that generates a new space where we can cut in.

The inset image on the right shows examples of visual semantics and its mask shape associated with actions (left, straight, and right) computed by ActionTunnel. Our method embeds such semantics into the present first person image to synthesize a transitional image.

6. Summary

This paper presents a method to synthesize a present first person image to afford the desired future action in a spatially consistent way. We leverage ActionTunnel to edit the present image to embed visual semantics of future action without manually annotation. We retrieve an image given the present scene based on similarity of future trajectories and 3D spatial layout. ActionTunnels from two images are constructed, segmented, glued, and projected onto the first person camera. The resulting image includes missing pixels where we complete using a generative adversarial network. We demonstrate that our method outperforms existing 2D based image inpainting and produces compelling visual experience by traveling through space and time.
Table 1. Normalized cross correlation (NCC) over frame difference: median(std.)

| Δt | 2   | 4   | 6   | 8   | 10  | 2   | 4   | 6   | 8   | 10  |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
|     | Indoor |     |     |     |     | Outdoor |     |     |     |     |
| 2D pasting | 0.35(0.19) | 0.40(0.18) | 0.35(0.20) | 0.30(0.19) | 0.25(0.19) | 0.48(0.23) | 0.40(0.24) | 0.35(0.25) | 0.31(0.27) | 0.30(0.24) |
| 2D GAN | 0.20(0.28) | 0.17(0.28) | 0.18(0.25) | 0.17(0.25) | 0.13(0.25) | 0.15(0.15) | 0.15(0.15) | 0.16(0.15) | 0.16(0.15) | 0.16(0.15) |
| 3D pasting | 0.39(0.18) | 0.39(0.18) | 0.39(0.20) | 0.39(0.18) | 0.28(0.19) | 0.50(0.24) | 0.42(0.26) | 0.36(0.26) | 0.32(0.25) | 0.31(0.24) |
| 3D GAN | 0.29(0.14) | 0.27(0.14) | 0.28(0.15) | 0.28(0.15) | 0.27(0.13) | 0.39(0.17) | 0.40(0.17) | 0.40(0.17) | 0.42(0.17) | 0.42(0.17) |
| 3D Box+GAN | 0.48(0.14) | 0.43(0.15) | 0.37(0.16) | 0.35(0.14) | 0.31(0.14) | 0.53(0.19) | 0.49(0.20) | 0.46(0.19) | 0.45(0.19) | 0.44(0.18) |
| Ours | 0.54(0.12) | 0.44(0.13) | 0.39(0.13) | 0.37(0.13) | 0.34(0.14) | 0.60(0.18) | 0.57(0.19) | 0.51(0.19) | 0.50(0.19) | 0.47(0.18) |

Figure 8. We apply our method to edit the present image such that it embeds the future desired path for diverse outdoor and indoor scenes. The second column to sixth column are synthesized images over time from present scene (left) to future scene (right). The color of the trajectory encodes the transitional point between scenes. Results are best seen in color.
References

[1] P. Agrawal, J. Carreira, and J. Malik. Learning to see by moving. 2015.
[2] I. Arev, H. S. Park, Y. Sheikh, J. K. Hodgins, and A. Shamir. Automatic editing of footage from multiple social cameras. SIGGRAPH, 2014.
[3] G. Bertasius, S. X. Yu, H. S. Park, and J. Shi. First person action-object detection with egonet. In arXiv:1603.04908, 2016.
[4] T. F. Brady, T. Konkle, and G. A. Alvarez. A review of visual memory capacity: Beyond individual items and toward structured representations. Journal of Vision, 2011.
[5] T. Dekel Basha, Y. Moses, and S. Avidan. Photo sequencing. In ECCV, 2012.
[6] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros. What makes paris look like paris? SIGGRAPH, 2012.
[7] A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. In ICCV, 2011.
[8] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interaction: A first-person perspective. In CVPR, 2012.
[9] J. J. Gibson. The Ecological Approach to Visual Perception. Boston: Houghton Mifflin, 1979.
[10] I. J. Goodfellow, J. Pouget-Abadie, B. Mirza, M. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. In arXiv:1406.2661, 2014.
[11] E. T. Hall. A system for the notation of proxemic behaviour. American Anthropologist, 1963.
[12] R. Hartley and A. Zisserman. Multiple View Geometry in Computer Vision. Cambridge University Press, second edition, 2004.
[13] J. Hernandez, Y. Li, J. M. Rehg, and R. W. Picard. Physiological parameter estimation using a head-mounted wearable device. In MOBILE HEALTH, 2014.
[14] Y. Horry, K.-I. Anjo, and K. Arai. Tour into the picture: Using a spidery mesh interface to make animation from a single image. SIGGRAPH, 1997.
[15] D. Jayaraman and K. Grauman. Learning image representations tied to ego-motion. In ICCV, 2015.
[16] I. Kemelmacher-Shlizerman, A. Sankar, E. Shechtman, and S. M. Seitz. Being john malkovich. In ECCV, 2010.
[17] I. Kemelmacher-Shlizerman, E. Shechtman, R. Garg, and S. M. Seitz. Exploring photosio. SIGGRAPH, 2011.
[18] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In CVPR, 2011.
[19] J. Kopf, M. Cohen, and R. Szeliski. First person hyperlapse videos. SIGGRAPH, 2014.
[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, 2012.
[21] Y. J. Lee, A. A. Efros, and M. Hebert. Style-aware mid-level representation with deep convolutional neural networks. In NIPS, 2012.
[22] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people in egocentric videos. In CVPR, 2013.
[23] Y. Li, J. Ghosh, and K. Grauman. Story-driven summarization for egocentric video. In CVPR, 2013.
[24] M. Ma, H. Fan, and K. M. Kitani. Going deeper into first-person activity recognition. In CVPR, 2016.