Associating Objects and Their Effects in Video through Coordination Games

Erika Lu[1] **Forrester Cole**[1] **Weidi Xie**[2] **Tali Dekel**[1,3]

William T. Freeman[1] **Andrew Zisserman**[4] **Michael Rubinstein**[1]

[1] Google Research [2] Shanghai Jiao Tong University [3] Weizmann Institute [4] University of Oxford

Abstract

We explore a feed-forward approach for decomposing a video into layers, where each layer contains an object of interest along with its associated shadows, reflections, and other visual effects. This problem is challenging since associated effects vary widely with the 3D geometry and lighting conditions in the scene, and ground-truth labels for visual effects are difficult (and in some cases impractical) to collect. We take a self-supervised approach and train a neural network to produce a foreground image and alpha matte from a rough object segmentation mask under a reconstruction and sparsity loss. Under reconstruction loss, the layer decomposition problem is underdetermined: many combinations of layers may reconstruct the input video. Inspired by the game theory concept of focal points—or *Schelling points*—we pose the problem as a coordination game, where each player (network) predicts the effects for a single object without knowledge of the other players' choices. The players learn to converge on the "natural" layer decomposition in order to maximize the likelihood of their choices aligning with the other players'. We train the network to play this game with itself, and show how to design the rules of this game so that the focal point lies at the correct layer decomposition. We demonstrate feed-forward results on a challenging synthetic dataset, then show that pretraining on this dataset significantly reduces optimization time for real videos.

1 Introduction

Identifying and associating effects such as shadows and reflections with the objects that produce them is a fundamental and difficult task for visual understanding. A variety of cues such as motion, appearance, and proximity may be used to match an object and an effect, but these cues may vary widely depending on the 3D geometry and lighting conditions of the scene, and nearby objects may produce similar effects (Fig 1). There is no simple heuristic to reliably connect an object and the visual effects it generates.

The recent *Omnimatte* approach [11, 12] learns to associate an object with its effects by training a CNN to decompose the video into layers, where each layer contains the object as well as the effects associated with that object. The CNN is given a binary segmentation mask of an object, omitting any associated effects, and learns to output a foreground color and alpha matte for that object and its effects. Notably, no explicit loss is provided for the association of objects and effects: the correct association emerges through the structure of the optimization and the inductive bias of the CNN. The optimization requires hours of processing for each video, however, and is vulnerable to errors when multiple objects have closely correlated motion (Fig 5).

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Figure 1: Example videos and our decomposition result with shadows and reflections. Matching objects and their effects is challenging: the true object may be further from its shadow (a) or reflection (b) than another object, and effects may be subtle (reflection (c), soft shadow (d)).

This paper aims to solve the layer decomposition problem directly by training a feed-forward neural network to predict the layers from the input video and masks. Since ground-truth labels for effects are not available for real videos, we take a self-supervised approach and train the network only with a reconstruction and sparsity loss. Unfortunately, reconstruction and sparsity losses do not uniquely determine a layer decomposition: effects may be shuffled between layers while producing the same loss (Fig 3). For example, a valid but incorrect solution puts the entire video in the front layer, leaving the remaining layers blank.

To solve this problem, we take inspiration from the game theory concept of focal points, also known as Schelling points [16]. In a coordination game, the players must cooperate to achieve a goal without communicating with each other ahead of time. A classic example is giving several people the instruction: "Meet in New York City on Tuesday", without giving any details about the exact location and time when they should meet. Schelling found that the most common strategy, based on the players' own prior knowledge of the city, was to go to Grand Central station at noon of that day. Players converge on such a "natural" solution, or focal point, in order to maximize the likelihood of their choices aligning with those of other players. In our case, the network plays a coordination game against itself: it is given an object mask and is asked to predict the appearance of that object and its visual effects, without knowledge of the other network instances' choices. We show that by training the network to predict masked video frames, we can engineer the focal point of this game to lie at the correct layer decomposition.

We demonstrate the success of the approach on a challenging synthetic dataset of moving objects with realistic shadows and reflections. The learned prior helps the network overcome ambiguous inputs, leading to improved quality over single-video optimization [12]. The prior learned from synthetic data also provides a good starting point for optimization on real videos. Fine-tuning the pretrained network on a new video reduces optimization time by 10 over single-video optimization, while producing similar layer decomposition quality.

2 Related Work

Video layer decomposition. Separating videos into layers is a fundamental computer vision problem that has been studied for decades [18]. It is important for both basic understanding of scenes (e.g., for estimating depth [24] and occlusion boundaries [2]), as well as for supporting various operations on videos such as editing (e.g., [1]) and view synthesis (e.g., [17]). Recently, several works besides the previously mentioned Omnimatte [11, 12] have applied deep learning to make progress on this problem. Zhang et al. [23] generate editable free-viewpoint videos of dynamic scenes from sequences captured by multiple cameras through a learned, layered neural representation that disentangles location, deformation, and appearance of dynamic objects. Kasten et al. [9] propose a decomposition method that represents a video as a set of layered 2D atlases that are global to the video, along with associated (per-frame) alpha maps. Like our work, these methods are self-supervised. However, they involve slow optimizations, overfitting neural networks' weights to an input video. In contrast, this work aims to solve the decomposition problem directly by training a feed-forward model to predict the layers from the input video and masks.