We thank the reviewers for their time and their reviews. We address the questions below.

Ambiguities beyond flips and rotations (R3) As pointed out by R3, the albedo of the hidden scene is fundamentally ambiguous, as any intensity can be compensated by a reciprocal intensity in the transport matrix. To anchor the solution colors, we use the common "gray world assumption", and impose it by a simple chromaticity prior that discourages large differences between color channels. The color of the observed scene can therefore tint the colors of the hidden scene solution.

The space of ambiguities and potential distortions can be characterized as follows (see Koenderink et al., The Generic Bilinear Calibration-Estimation Problem). Let T_0 and L_0 be the true underlying factors, the observed video thus being $Z = T_0 L_0$. All "valid" factorizations are of the form $T = T_0 A^\dagger$ and $L = A L_0$, where A is chosen (almost) arbitrarily and $A^\dagger$ is its (pseudo)inverse. This can be seen by substituting $TL = (T_0 A^\dagger)(A L_0) = T_0 (A^\dagger A) L_0 = T_0 L_0 = Z$.

The result of any factorization implicitly corresponds to some choice of A and $A^\dagger$. In simple cases, the matrix A can represent e.g. a permutation that flips the image, whence $A^\dagger$ is a flip that restores the original orientation: this case is illustrated in Figure 4's conversely flipped matrices. They can also represent complementary color transformations as discussed above. However, for classical factorization methods, they tend to consist of unstructured "noise" that scrambles the image-like structure in T_0 and L_0 beyond recognition.

Our finding in the paper is that via DIP-based factorization, these transformations instead tend to express continuous and bijective image warps (and color modulations) that preserve the general image structure. As observed by R3, this does in practice include more complex distortions than just flips and rotations —- see for example the nonlinear stretching of the cameraman image in Figure 4. In full two dimensions, there is room for more complex distortions, but we still find that e.g. the relative motions of independent objects often remain readable.

Geometric complexity (R3) We assume that the scene contains a sufficient amount of geometric complexity to generate high-frequency features like shadows. This improves the conditioning of the problem, as discussed in the literature on frequency analysis of light transport effects (see e.g. A Theory of Locally Linear Light Transport by Mahajan et al.). We will emphasize this in the revised paper.

Comparisons to previous methods (R4) To our knowledge, no existing work attempts to solve the problem under a similarly general setup, with no assumptions about the shapes viewed in the scene. Attempts to use standard factorization methods consistently produce unstructured and scrambled results, analogous to the baselines in Figure 4. An example is seen in the supplemental video, where an SVD factorization is visualized at time 2:00 - 2:09.

To provide a comparison, we generalized a recent algorithm that addresses the closest analogue we could think of, i.e. blind deconvolution with a classical sparse gradient prior that models natural image statistics. As discussed in Section 5.2, we were unable to obtain competitive results despite fair efforts put into the experiment.

Regarding comparisons to other non-line-of-sight methods: active non-line-of-sight methods are outside the scope of the paper, as these techniques assume fundamentally different imaging modalities (usually static hidden scenes, actively probed over an extended period of time). Similarly, a recent passive non-line-of-sight technique by Bouman et al. (Turning Corners into Cameras: Principles and Methods, ICCV 2017) assumes a specific scene geometry with clearly defined "corners" and focuses on near-invisible signals, while our method assumes the geometry and reflectances of the relay objects are unknown.

Validity of reconstructions for machine vision tasks (R4) When factorizing light transport using traditional factorization techniques, there is no guarantee that the two factors correspond to the true visible and hidden scene, or even to any plausible image signal. The question raised by R4 is whether the factors reconstructed using DIP actually correspond to the true visible and hidden scenes, or are just arbitrary natural-looking images. Figure 6 of the paper provides a partial answer to that question. We show that for controlled scenes such as the Disks sequence, the reconstructed signal is clearly not arbitrary, but closely matches the ground truth sequence. Even for more complicated sequences (Hands), this correspondence still appears to hold.