Neural Image Representations for Multi-Image Fusion and Layer Separation

Seonghyeon Nam, Marcus A. Brubaker, and Michael S. Brown

York University
snam0331@gmail.com, {mab,mbrown}@eecs.yorku.ca

Abstract. We propose a framework for aligning and fusing multiple images into a single view using neural image representations (NIRs), also known as implicit or coordinate-based neural representations. Our framework targets burst images that exhibit camera ego motion and potential changes in the scene. We describe different strategies for alignment depending on the nature of the scene motion—namely, perspective planar (*i.e.*, homography), optical flow with minimal scene change, and optical flow with notable occlusion and disocclusion. With the neural image representation, our framework effectively combines multiple inputs into a single canonical view without the need for selecting one of the images as a reference frame. We demonstrate how to use this multi-frame fusion framework for various layer separation tasks. The code and results are available at https://shnnam.github.io/research/nir.

Keywords: Implicit neural representations, coordinate-based neural representations, multi-image fusion, layer separation

1 Introduction and Related Work

Fusing multiple misaligned images into a single view is a fundamental problem in computer vision. The underlying assumption for this task is that the multiple images represent varying viewpoints of the same scene, perhaps with small motion in the scene. Many computer vision tasks rely on multi-image fusion, such as image stitching [16,13,7], high dynamic range (HDR) imaging [11,43,42], and image super-resolution [6,5,39]. Most existing image fusion approaches work by first aligning the multiple images based on their assumed motion—for example, homography for planar or nearly planar scenes or optical flow for nonplanar scenes, or when objects in the scene move. Traditionally, images are aligned to a reference image that is manually chosen among the input images. Since image pixels are represented in a 2D discrete sampled array, such transformations are approximated by interpolation techniques.

Recently, implicit or coordinate-based neural representations were proposed to represent images and videos as a function of pixel coordinates parameterized by multi-layer perceptrons (MLPs) [36,32]. This new type of image representation, which we call a neural image representation (NIR), is different from

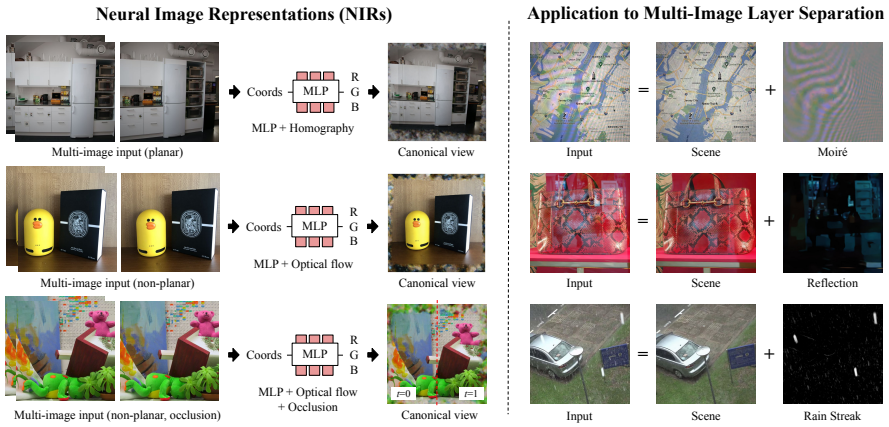


Fig. 1: This figure provides an overview of our work which fuses multiple images to a single canonical view in a continuous image representation. Our method incorporates motion models such as homography and optical flow into the formulation of implicit or coordinate-based neural representations. We demonstrate the effectiveness of our method on various applications of multi-image layer separation. Images from [4,14,8,23] are used here for visualization.

conventional discrete grid-based representations in that image signals are continuous with respect to spatial or spatio-temporal pixel coordinates. Further, the resolution of images no longer depends on the size of the discrete grid, but rather the representational complexity of the MLP. These representations have been actively studied particularly in view synthesis [25,33,30,15,22,27,28], 3D geometry [24,26], and image synthesis [34,2,31].

This work targets multi-frame fusion by leveraging the advantages offered by NIRs. As shown on the left in Fig. 1, we propose to train MLPs to reconstruct a canonical view based on multiple images. Our approach incorporates image registration techniques into NIRs using coordinate transformations [27,28]. Unlike existing multi-image fusion, our method does not need an explicit reference image. Instead, a virtual reference image is implicitly learned as the canonical view embedded within the neural representation. Since the space of canonical views is unbounded, all images can be fused regardless of the original image frame as shown in Fig. 1. In addition, image transformation is achieved in a real-valued coordinate space without the need for interpolation.

To demonstrate effectiveness of our NIR multi-image fusion, we apply our method to various applications of multi-image layer separation. As shown in Fig. 1, the goal of multi-image layer separation is to decompose signals from multiple images into a single underlying scene image and interference layers to improve the visibility of the underlying scene. Many approaches for different tasks have been studied, such as image demoiré [12,40], reflection removal [14,1,19,20], fence removal [19,20], and deraining [10,44,8,38]. Early works on these problems heav-

ily rely on domain-specific priors for optimization, while recent approaches are driven by deep learning and a large amount of annotated data for supervision.

In our work, we cast the problem as an unsupervised optimization of NIRs. Specifically, we fuse the underlying scene from the multiple images using NIRs. Depending on the type of scene motion, we use different deformation strategies when computing the neural image representation for each frame. To remove the interference layer, we propose two-stream NIRs. In particular, the underlying "clean layer" image without interference is represented by one MLP, while a separate MLP is used to represent the interference layer(s). We show that standard regularization terms – for example, total variation – can be used in the optimization of these NIRs to assist in the layer separation. We demonstrate the effectiveness of our approach on moiré removal, obstruction removal, and rain removal.

Closely related to our approach is DoubleDIP [9], which also studied image layer decomposition on a single or multiple image(s) using coupled deep image priors [37]. DoubleDIP exploits self-similarity, an inductive bias in convolutional neural networks (CNNs), to separate different signals. Our approach uses the parameterization of motion as a general prior to tackle multiple tasks. Unlike semantic segmentation and matting [9,21], we focus on disentangling low-level signals rather than semantic layers.

Contribution. We describe a framework to perform multi-frame fusion and layer separation as learning a neural image representation. We describe variations on the representation and optimization for different scene and camera conditions. We also demonstrate how to apply this framework to handle several different types of layer separation tasks. To the best of our knowledge, our work is the first to explicitly address multi-image fusion with neural image representations.

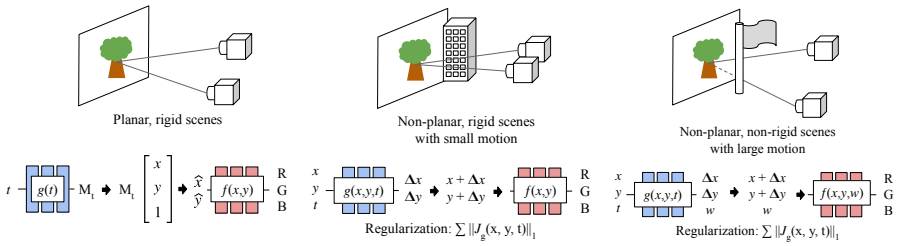
2 Method Overview

Neural image representations, also known as implicit or coordinate-based neural representations, have recently been proposed [32,36] as a way to represent RGB values of an image as a function of pixel coordinates parameterized by MLPs. For multiple sequential images, this can be formulated as

$$\hat{\mathbf{I}}_{(x,y,t)} = f_{\theta_I}(x, y, t), \tag{1}$$

where $\hat{\mathbf{I}}_{(x,y,t)}$ is the value at pixel (x, y) in frame t and f_{θ_I} is an MLP with parameters θ_I. Here each frame is nearly independent due to different values of t.

In our work, we assume our multiple images are captured quickly as a burst from a single camera. Consequently, images are of approximately the same scene, but are expected to have small variations due to the motion of the camera and small amounts of motion in the scene. Furthermore, we do not expect notable variations in scene lighting, appearance, or colors due to the camera's onboard color manipulation.



(a) Homography-based NIR (b) Occlusion-free flow-based NIR (c) Occlusion-aware flow-based NIR

Fig. 2: Illustration of our neural image representations (NIRs). Assuming that the MLP f learns a canonical view where all burst images are fused, we render each image by projecting the canonical view to the frame-specific view, which is achieved by transforming the input coordinates fed into the f. We estimate the transform using another MLP g. According to different assumptions of the world, we formulate our framework differently; we formulate the transform of coordinates using (a) homography, (b) optical flow without occlusion/disocclusion, and (c) optical flow with occlusion/disocclusion.

Within this context of burst images, our work aims to formulate f_{θ_I} differently by learning a joint representation of multiple images using their spatio-temporal correlation. To this end, we revisit well-established image registration and motion compensation techniques within the new domain of NIRs. Specifically, our f_{θ_I} learns a canonical view of the scene shared across images. Each image in the burst sequence is modelled by a deformation of the canonical view—for instance, using a perspective planar transform (i.e., a homography) or pixel-wise optical flow. Since the function is continuous and unbounded, it not only is able store the entire scene regardless of the size of 2D image grid, but also can be easily deformed by transforming input coordinates into a real-valued space. The model is formally described as

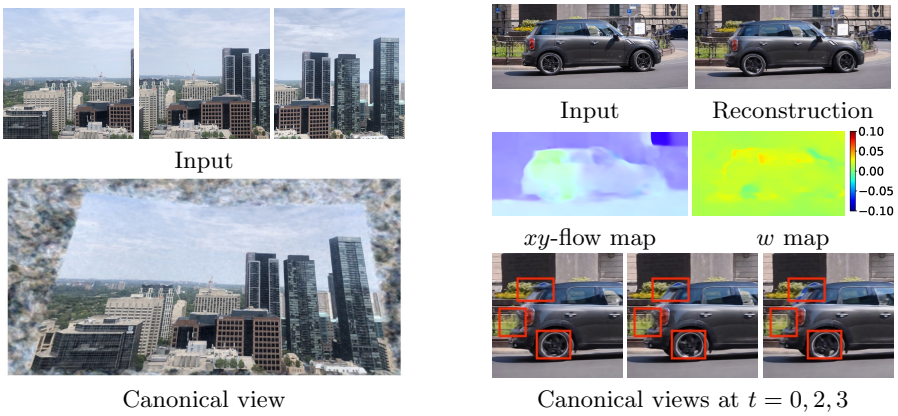
$$\hat{\mathbf{I}}_{(x,y,t)} = f_{\theta_I}(T_g(x, y, t)), \tag{2}$$

where T applies a coordinate transformation with parameters ϕ. The parameters of the coordinate transform could be fixed or themselves a function— that is, $g = g_{\theta_T}(x, y, t)$ where g_{θ_T} is an MLP that computes the parameters of the coordinate transform. The parameters of the MLPs, θ_T and θ_I, are optimized by minimizing the following pixel reconstruction loss:

$$\mathcal{L}_{\text{Recon}} = \sum_{x,y,t} \|\hat{\mathbf{I}}_{(x,y,t)} - \mathbf{I}_{(x,y,t)}\|_2^2, \tag{3}$$

where $\mathbf{I}$ is the original image ground truth.

The explicit parameterization of motion in neural representations enables the simultaneous learning of image and motion representations. By minimizing Eq. (3), our neural representations learn the parameters of scene motion



(a) Homography-based NIR (b) Occlusion-aware flow-based NIR

Fig. 3: Visualization of learned representations. In (a), the top row shows three of nine representative images used to learn a homography-based NIR, and the bottom shows a learned canonical view. In (b), the first row shows one of the input and reconstruction images, the second row shows a xy-flow map and w map learned by a occlusion-aware flow-based NIR, and the third row shows canonical views at $t = 0, 2, 3$.

in an unsupervised manner. More importantly, unlike conventional image registration and motion compensation techniques, our approach does not require a reference image to be selected from the burst input. Instead, our model learns a virtual reference view of the scene implicitly.

We next show how to extend our multi-frame alignment framework for use in layer separation tasks. In particular, we target tasks where input images are modeled as a combination of two layers: (1) the desired underlying scene image and (2) the undesired corruption in the form as an interference layer. We assume that the contents of the underlying scene remains similar over the multiple images, while the interference layer changes. To do this, we propose a two-stream architecture for NIRs, with one component that captures the static scene and another that captures the interference. In the following, we describe our method in detail.

2.1 NIRs for Multi-Image Fusion

Fig. 2 shows the overview of the NIRs for multi-image fusion. We propose three kinds of parameterization according to the assumption of the scene: (a) homography-based NIRs, (b) occlusion-free flow-based NIRs, and (c) occlusion-aware flow-based NIRs.

Homography-based NIRs. In case of planar, rigid scenes that are moving globally as shown in Fig. 2 (a), we can use a homography as the coordinate transforma-

tion. As shown in the figure, the function g_{θ_T} is learned to estimate parameters of a homography matrix M for each frame. Then the predicted image using the homography-based NIRs is described as:

$$\hat{\mathbf{I}}_{(x,y,t)} = f_{\theta_I}(M_t[x, y, 1]^T), \tag{4}$$

where M_t is a 3×3 linear matrix represented as $M_t = g_{\theta_T}(t)$. Since M_t is applied globally regardless of spatial coordinates, $g_{\theta_T}(t)$ only takes t as input. We omit the normalization of output coordinates in the homography transform for simplicity.

Fig. 3 (a) shows a visualization of the NIR estimated from nine burst images of a distant scene, captured with a horizontally moving camera. As can be seen in the figure, the homography-based NIR automatically stitches all the images in a single view only using a reconstruction loss. A single frame t can be recreated by transforming the canonical view using the output of the $g_{\theta_T}(t)$ homography matrix.

Occlusion-free flow-based NIRs. In many cases, a scene will not be planar or move together rigidly. However, in burst imagery because frames are temporally close, the motions are likely to be small. To handle this, we use a dense optical flow representation to model the per-pixel displacement of scene, which is represented by the displacement of x and y coordinates as shown in Fig. 2 (b). We assume that the motion is small enough to cause minimal occlusions and disocclusions. In this case ϕ represents an xy-displacement that is computed by $g_{\theta_T}(x, y, t)$ for each (x, y, t). Formally, $T(x, y, t) = (x + \Delta x_t, y + \Delta y_t)$ where $(\Delta x_t, \Delta y_t) = g_{\theta_T}(x, y, t)$ are the displacement of x and y coordinates. An output pixel can be computed as:

$$\hat{\mathbf{I}}_{(x,y,t)} = f_{\theta_I}(x + \Delta x_t, y + \Delta y_t). \tag{5}$$

In addition to the reconstruction loss in Eq. (3), we use a total variation (TV) regularization for the flow smoothness, which is described as

$$\mathcal{L}_{\text{TVFlow}} = \sum \|J_{g_{\theta_T}}(x, y, t)\|_1, \tag{6}$$

where $J_{g_{\theta_T}}(x, y, t)$ is a Jacobian matrix that consists of gradients of g_{θ_T} with respect to x, y, and t.

Occlusion-aware flow-based NIRs. Since the canonical view of the occlusion-free flow-based NIRs is in a 2D plane, it is not enough to store extra information when a scene is occluded or disoccluded. To address such cases, we add an additional dimension w to the canonical view as shown in Fig. 2 (c). Intuitively, different versions of a scene at a certain position caused by occlusion are stored at different values of w, while occlusion-irrelevant pixels are stored at the same value of w and shared across images. This is achieved by regularizing the Jacobian of g_{θ_T} in Eq. (6). With w, the output image is rendered by the following equation:

$$\hat{\mathbf{I}}_{(x,y,t)} = f_{\theta_I}(x + \Delta x_t, y + \Delta y_t, w_t). \tag{7}$$

Fig. 3 (b) shows a visualization of a learned xy-flow map, w map, and canonical views at different values of w after training five consecutive images in [29]. Since the car is moving in the scene, the xy-flow map shows spatially varying optical flow on the car. The w map shows different values in regions of large motion (e.g., wheels), transient lighting effects (e.g., specularities and reflections), and regions that undergo occlusion or disocclusion. This can be seen more clearly by visualizing the canonical view, with different values of w as shown in the bottom of the figure.

2.2 Two-Stream NIRs for Layer Separation

We now extend NIRs to multi-image layer separation tasks. Fig. 4 shows the overview of our two-stream NIRs. We model the images as the combination of two signals,

$$\begin{cases} \hat{\mathbf{O}}_{(x,y,t)} & = f^1_{\theta_O}(T_g(x,y,t)), \\ \hat{\mathbf{U}}_{(x,y,t)} & = f^2_{\theta_U}(x,y,t) \end{cases} \tag{8}$$

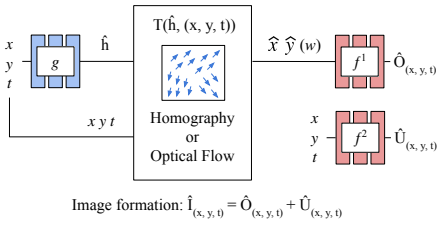


Fig. 4: Multi-image layer separation.

where $f^1_{\theta_O}$ and $f^2_{\theta_U}$ are two different MLPs used to represent the scene and corrupting interference, respectively.

Since we usually have the knowledge of scene motion, we use an explicit parameterization of motion for $f^1_{\theta_O}$—for example, a homography or a flow field. To model the interference layers, we use an unconstrained form of MLP for $f^2_{\theta_U}$ to store contents that violate the motion in $f^1_{\theta_O}$, that is beneficial for interference patterns difficult to model. The generic form of image formation is described as

$$\hat{\mathbf{I}}_{(x,y,t)} = \hat{\mathbf{O}}_{(x,y,t)} + \hat{\mathbf{U}}_{(x,y,t)}, \tag{9}$$

but the specifics can vary depending on the task. Due to the flexibility of $f^2_{\theta_U}$, it can potentially learn the full contents of the images as a "video", effectively ignoring $f^1_{\theta_O}$. To prevent this, we regularize $f^2_{\theta_U}$ using

$$\mathcal{L}_{\text{Interf}} = \sum \|\hat{\mathbf{U}}_{(x,y,t)}\|_1. \tag{10}$$

Directly incorporating a spatial alignment into the NIR optimization may appear inefficient at first glance, especially compared to methods that first apply conventional homography and or optical flow estimation and then perform some type of image fusion. However, in the case of corrupted scenes, it is often challenging to estimate the motion of the underlying clean image with the conventional methods. For instance, existing methods often rely heavily on multiple stages of refinement of motion [19] to pre-process images to assist with the alignment step. Our method tackles the problem jointly, by incorporating the scene alignment jointly with a layer separation through the benefits of NIRs.

3 Applications

We now show the effectiveness of our method on various multi-image layer separation tasks. Please refer to the supplementary material for more results.

3.1 Moiré Removal

Moiré is a common pattern of interference, often seen when taking a photo of monitor or screen using a digital camera. Moiré patterns are caused by the misalignment of the pixel grids in the display and camera sensor. Burst images usually capture temporally varying moiré patterns as camera motion changes the alignment of the sensor and screen and hence the interference pattern. Typically the movement of the scene in burst images follows homography transform as the screen is planar. We show that our two-stream NIRs are able to effectively separate the underlying scene and moiré pattern.

Formulation. We parameterize $f^1_{\theta_O}$ as a homography-based NIR in Eq. (4). The image formation follows the basic form in Eq. (9), where we use signed values in the range of $[-1, 1)$ for the output of both $\hat{\mathbf{O}}_{(x,y,t)} \in \mathbb{R}^3$ and $\hat{\mathbf{U}}_{(x,y,t)} \in \mathbb{R}^3$. The signed output for $\hat{\mathbf{U}}_{(x,y,t)}$ is particularly useful to represent color bands of moiré patterns. To further prevent scene content from appearing in both $\hat{\mathbf{O}}_{(x,y,t)}$ and $\hat{\mathbf{U}}_{(x,y,t)}$, we adopt an exclusion loss used in [9,45] to encourage the gradient structure of two signals to be decorrelated. This is formulated as

$$\mathcal{L}_{\text{Excl}} = \sum \|\Phi(J_{f^1}(x,y), J_{f^2}(x,y))\|_2^2, \tag{11}$$

where $\Phi(J_{f^1}(x,y), J_{f^2}(x,y)) = \tanh(N_1 J_{f^1}(x,y)) \otimes \tanh(N_2 J_{f^2}(x,y))$, and $\otimes$ is an element-wise multiplication. N_1 and N_2 are normalization terms [45]. We optimize θ_T, θ_O, and θ_U using the following training objective:

$$\mathcal{L}_{\text{Moire}} = \mathcal{L}_{\text{Recon}} + \lambda_{\text{Interf}}\mathcal{L}_{\text{Interf}} + \lambda_{\text{Excl}}\mathcal{L}_{\text{Excl}}, \tag{12}$$

where λ_{Interf} and λ_{Excl} are hyperparameters. We use an MLP with ReLU activation for g_{θ_T} and a SIREN [32] for $f^1_{\theta_O}$ and $f^2_{\theta_U}$.

Experiments. Since there are no publicly available datasets for multi-frame screen-captured moiré images, we synthesize a dataset from clean images. To do this, we use the Slideshare-1M [3] dataset, which consists of approximately one million images of lecture slides, to mimic content likely to be captured by students. Using this dataset, we synthesize 100 test sequences of five burst images for testing following the synthesis procedure in [17]. For comparison, we compare AFN [40] and C3Net [12], state-of-the-art deep learning methods, which are trained by our synthetic training set containing 10,000 sequences. We additionally evaluate Double DIP to compare unsupervised approaches.

Table 1 shows a quantitative comparison of methods on the synthetic test set. In addition to PSNR and SSIM, we compare a normalized cross-correlation

	Supervised		Unsupervised	
	AFN [40]	C3Net [12]	Double DIP [9]	Ours
Input	Single	Burst	Burst	Burst
PSNR	43.63	27.99	18.53	38.68
SSIM	0.9952	0.8071	0.8762	0.9751
NCC	0.9963	0.7724	0.5120	0.9865
SI	0.9962	0.7721	0.4895	0.9856

Table 1: Quantitative evaluation of moiré removal with a synthetic dataset. AFN [40] uses a single image; other methods use five images as input.

(NCC) and structure index (SI). Even though our method does not outperform AFN, the performance is significantly better than C3Net and Double DIP. However, notably our method is unsupervised—that is, it does not use a training set of images. This is in contrast to AFN and C3Net, which require explicit supervision or clean and moiré corrupted images. Fig. 5 shows a qualitative evaluation on real images. As can be seen, our method outperforms all the baselines on real images. The performance of AFN and C3Net is degraded because they are not trained on real images. Double DIP fails to decompose the underlying scene and moiré pattern since it relies on an inductive bias in convolutional neural networks, which is not enough to separate complex signals. Our method removes a moiré pattern by restricting the movement of the scene to homography, which acts as a strong prior of moiré removal.

3.2 Obstruction Removal

The goal of obstruction removal [20,41] is to eliminate foreground objects or scenes that hinder the visibility of the background scene. Obstructions can be in the form of reflection on a window in front of the scene or a physical object, such as a fence. We apply the two-stream NIRs based on occlusion-free optical flow to a reflection and fence removal. In this case, the background scenes are not planar, but the movement of the scene is small enough to ignore occlusion. Similarly to moiré removal, we decompose the reflection and fence layer using the fact that they move differently to the background scene.

Formulation. We use the occlusion-free flow-based NIRs in Eq. (5) for $f^1_{\theta_O}$. For reflection removal, we use the image model in Eq. (9), where $\hat{\mathbf{O}}_{(x,y,t)} \in \mathbb{R}^3$ and $\hat{\mathbf{U}}_{(x,y,t)} \in \mathbb{R}^3$ are in the range of $[0, 1)$. We use the following combination of loss functions as a training objective:

$$\begin{aligned}\mathcal{L}_{\text{Refl}} =& \mathcal{L}_{\text{Recon}} + \lambda_{\text{TVFlow}}\mathcal{L}_{\text{TVFlow}} \\ &+ \lambda_{\text{Interf}}\mathcal{L}_{\text{Interf}} + \lambda_{\text{Excl}}\mathcal{L}_{\text{Excl}},\end{aligned} \tag{13}$$

where λ_{TVFlow} is a hyperparameter. For a fence removal, we use a different image model described as

$$\hat{\mathbf{I}}_{(x,y,t)} = (1 - \alpha_{(x,y,t)})\hat{\mathbf{O}}_{(x,y,t)} + \alpha_{(x,y,t)}\hat{\mathbf{U}}_{(x,y,t)}, \tag{14}$$

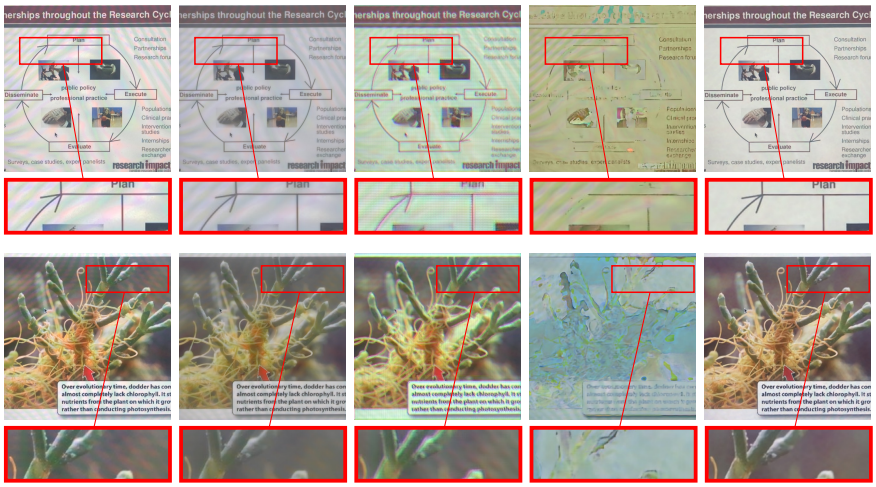


Fig. 5: Qualitative comparison of moiré removal on real images. Our method outperforms all methods including AFN [40]. AFN was better than ours on the synthetic data in Table 1 which is unrepresentative of real-world images.

where $(\alpha_{(x,y,t)}, \hat{\mathbf{U}}_{(x,y,t)}) = f^2_{\theta_U}(x,y,t)$, and $\hat{\mathbf{O}}_{(x,y,t)} \in \mathbb{R}^3$ and $\hat{\mathbf{U}}_{(x,y,t)} \in \mathbb{R}^3$ are in the range of $[0,1)$. $\alpha_{(x,y,t)} \in \mathbb{R}$ is an alpha map of the fence layer in the range of $[0,1)$. The training objective is described as:

$$\mathcal{L}_{\text{Fence}} = \mathcal{L}_{\text{Recon}} + \lambda_{\text{TVFlow}}\mathcal{L}_{\text{TVFlow}} + \lambda_{\text{Interf}}\mathcal{L}_{\text{Interf}}. \tag{15}$$

We used SIREN for all coordinate functions.

Experiments. Fig. 6 shows qualitative results of our method and existing approaches. We use real images in [14] for testing. The methods of Li and Brown [14] and Alayrac et al. [1] are designed for reflection removal, and the method in [20] is a general approach for obstruction removal. As can be seen, our method is able to accurately decompose the background scene and reflection compared with the baseline methods. Fig. 7 shows a qualitative comparison of fence removal on real images in [20]. Our method achieves comparable quality of results to learning-based methods that heavily rely on large amounts of data and supervision.

3.3 Rain Removal

To show the effectiveness of the occlusion-aware flow-based NIRs, we address the problem of multi-image rain removal as the task deals with various kinds of scenes, from static scenes to dynamic scenes. Since rain streaks move fast and randomly, the streaks impact the smoothness of the scene motion. We exploit this prior knowledge of the randomness of rain streaks observed in multiple images by imposing a smoothness regularization on the scene flow map.

Fig. 6: Qualitative results of reflection removal on real images in [14].

Fig. 7: Qualitative comparison of fence removal on real images in [20].

Formulation. We use the occlusion-aware flow-based NIRs in Eq. (7) as a formulation of $f^1_{\theta_O}$. Since rain streaks are achromatic, we use the following image formation:

$$\hat{\mathbf{I}}_{(x,y,t)} = (1 - \hat{\mathrm{U}}_{(x,y,t)})\hat{\mathbf{O}}_{(x,y,t)} + \hat{\mathrm{U}}_{(x,y,t)}\mathbf{1}, \tag{16}$$

where $\hat{\mathbf{O}}_{(x,y,t)} \in \mathbb{R}^3$ and $\hat{\mathrm{U}}_{(x,y,t)} \in \mathbb{R}$ are in the range of $[0, 1)$, and $\mathbf{1} = [1, 1, 1]^T$. In this form, $\hat{\mathrm{U}}_{(x,y,t)}$ acts as an alpha map of rain streaks. Our final training objective is described as

$$\mathcal{L}_{\text{Rain}} = \mathcal{L}_{\text{Recon}} + \lambda_{\text{TVFlow}}\mathcal{L}_{\text{TVFlow}} + \lambda_{\text{Interf}}\mathcal{L}_{\text{Interf}}. \tag{17}$$

Experiments. Fig. 8 shows a qualitative evaluation on real images in NTU-Rain [8], with moving cars and pedestrians. We compare state-of-the-art video deraining methods based on optimization [10] and deep learning [8]. We take five consecutive images to run our method which clearly removes rain streaks in the scene and is qualitatively competitive with the baselines on real images. For quantitative evaluation, we must resort to synthetic data and use RainSynLight25 [18], consisting of 25 synthetic sequences of nine images. Table 2 shows results of ours and baseline methods: SE [38], FastDeRain [10], SpacCNN [8], and

	Supervised		Unsupervised		
	SpacCNN [8]	FCDN [44]	SE [38]	FastDeRain [10]	Ours
PSNR	32.78	35.80	26.56	29.42	28.61
SSIM	0.9239	0.9622	0.8006	0.8683	0.8604

Table 2: Result of rain removal on RainSynLight25 [18].

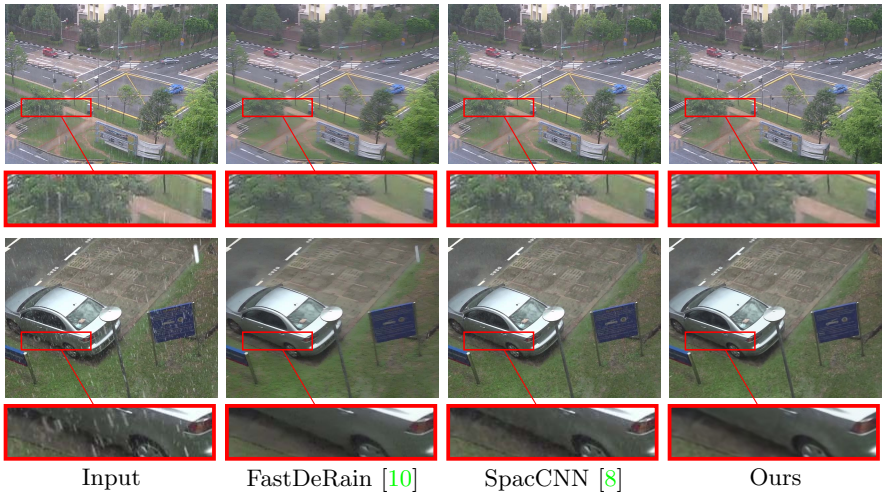


Fig. 8: Qualitative comparison of rain removal on real images in NTURain [8].

FCDN [44]. Though our method does not outperform deep learning-based methods, it achieves a comparable result to optimization-based approaches without supervision and the domain knowledge of deraining. We expect incorporating more regularization could further improve the performance.

3.4 Discussion

Ablation study on loss functions. We conducted an ablation study on various loss functions. In Fig. 9, we show the decomposed background and reflection layer of different training objectives by removing each loss function. As can be seen, the background content is reconstructed in the reflection layer when we do not use $\mathcal{L}_{\text{Interf}}$ since g_{θ_U} is unconstrained. Without $\mathcal{L}_{\text{TVFlow}}$, on the other hand, both signals are reconstructed in the background layer. In this case, g_{θ_O} has enough freedom to learn the mixture of two layers moving differently. In addition to $\mathcal{L}_{\text{Interf}}$ and $\mathcal{L}_{\text{TVFlow}}$, the exclusion loss $\mathcal{L}_{\text{Excl}}$ further improves the quality by preventing the structure of two layers from being correlated.

Ablation study on w. Fig. 10 shows an ablation study on w in the occlusion-aware flow-based NIRs using RainSynLight25 [18]. As shown in the red boxes

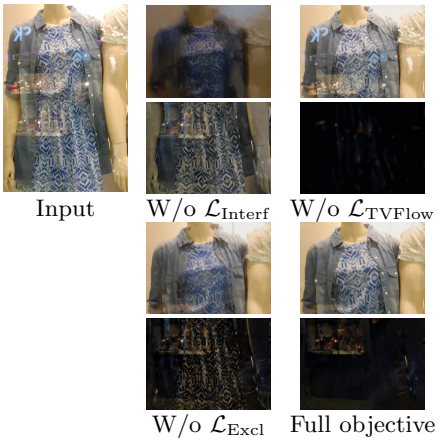


Fig. 9: Ablation study of loss functions on reflection removal. The top and bottom images show the background and reflection layer, respectively.

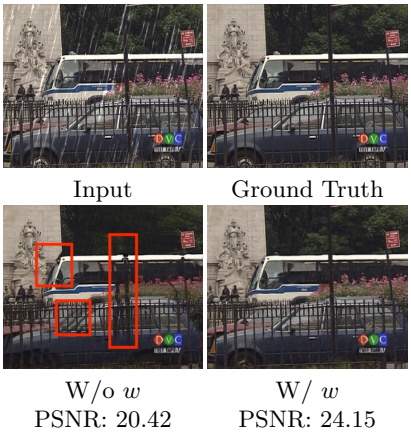


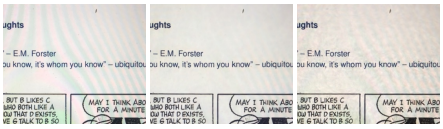
Fig. 10: Ablation study on w. We show PSNRs of two outputs using the synthetic dataset RainSynLight25 [18].

on the output, the method without w produces artifacts around occlusion and disocclusion, which indicates that it is difficult to represent all contents including occlusion in a 2D canonical view. Our method stores occluded appearance information in the extra dimension w, and enables accurate reconstruction.

Can a complex model take place of a simpler model? Note that in principle our NIRs using a complex motion model (e.g. occlusion-aware flow) can be generalized to the scenes with a simpler motion. For layer separation, however, it is beneficial to use a simpler model that fits well with the motion of a scene as it provides a strong constraint to separate layers effectively. In Fig. 11, we compare the homography-based model and the occlusion-aware flow-based model on a demoiréing task. The PSNR and SSIM of the flow-based model on the synthetic test set are 36.72 and 0.9512, respectively. The flow-based model removes the moiré pattern to some extent, but a part of the pattern still remains. This is because constraining the representation of motion in the homography-based model is more effective than adding regularization losses in the flow-based model.

The number of input images. Fig. 12 shows results of rain removal using the different number of input images. Better results are obtained with more images, as the additional images provide more information in separating two layers.

Limitations. Since our method relies on a pixel distance loss to learn the motion, it may fail when the motion of burst images is too large. Our method also fails to separate layers when the underlying scene and interference move in a similar manner. Although our method is not the top performer in all cases, it achieves



Input Homography Optical flow

Fig. 11: Analysis on different motion models. We apply homography-based (center) and occlusion-aware flow-based (right) models to demoiréing.

Input 2 images 5 images

Fig. 12: Analysis on the number of input images. We test 2 and 5 input images on a rain removal task.

competitive results without the need for supervision, which is the case for many of the state-of-the-art methods.

In addition, our method requires a proper assumption of motion to tackle layer separation tasks. This is because our method relies on the motion of underlying scene as a prior to separate layers. It may be more desirable to seek a generic model that works for any example without the assumption of motion by incorporating other priors such as an inductive bias learned from a large dataset.

Finally, our method currently takes about 30 minutes at most for optimizing layer separation tasks, which is a limiting factor in a real-world setting. However, there is already promising research demonstrating how to improve the optimization performance of NIRs [35].

4 Conclusion

We presented a framework that uses neural image representations to fuse information from multiple images. The framework simultaneously registers the images and fuses them into a single continuous image representation. We outlined multiple variations based on the underlying scene motion: homography-based, occlusion-free optical flow, and occlusion-aware optical flow. Unlike conventional image alignment and fusion, our approach does not need to select one of the input images as a reference frame. We showed our framework can be used to address layer separation problems using two NIRs, one for the desired scene layer and the other for the interference layer.

Neural image representations are an exciting new approach to image processing. This work is a first attempt to extend NIRs to multi-frame inputs with applications to various low-level computer vision tasks. Despite making only minimal assumptions and without leveraging any supervisory training data, the NIR-based approaches described here are competitive with state-of-the-art, unsupervised methods on individual tasks. Further, because it is practically impossible to acquire supervisory data in real-world conditions, our approach often qualitatively outperforms supervised methods on real-world imagery.

Acknowledgement. This work was funded in part by the Canada First Research Excellence Fund (CFREF) for the Vision: Science to Applications (VISTA) program and the NSERC Discovery Grant program.

References

1. Alayrac, J.B., Carreira, J., Zisserman, A.: The visual centrifuge: Model-free layered video representations. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2457–2466 (2019) 2, 10, 11
2. Anokhin, I., Demochkin, K., Khakhulin, T., Sterkin, G., Lempitsky, V., Korzhenkov, D.: Image generators with conditionally-independent pixel synthesis. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 14278–14287 (2021) 2
3. Araujo, A., Chaves, J., Lakshman, H., Angst, R., Girod, B.: Large-scale query-by-image video retrieval using bloom filters. arXiv preprint arXiv:1604.07939 (2016) 8
4. Baker, S., Scharstein, D., Lewis, J., Roth, S., Black, M.J., Szeliski, R.: A database and evaluation methodology for optical flow. Int. J. Comput. Vis. **92**(1), 1–31 (2011) 2
5. Bhat, G., Danelljan, M., Timofte, R.: Ntire 2021 challenge on burst super-resolution: Methods and results. In: IEEE Conf. Comput. Vis. Pattern Recog. Worksh. pp. 613–626 (2021) 1
6. Bhat, G., Danelljan, M., Van Gool, L., Timofte, R.: Deep burst super-resolution. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 9209–9218 (2021) 1
7. Brown, M., Lowe, D.G.: Automatic panoramic image stitching using invariant features. Int. J. Comput. Vis. **74**(1), 59–73 (2007) 1
8. Chen, J., Tan, C.H., Hou, J., Chau, L.P., Li, H.: Robust video content alignment and compensation for rain removal in a cnn framework. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 6286–6295 (2018) 2, 11, 12
9. Gandelsman, Y., Shocher, A., Irani, M.: Double-DIP: Unsupervised image decomposition via coupled deep-image-priors. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 11026–11035 (2019) 3, 8, 9, 10, 11
10. Jiang, T.X., Huang, T.Z., Zhao, X.L., Deng, L.J., Wang, Y.: Fastderain: A novel video rain streak removal method using directional gradient priors. IEEE Trans. Image Process. **28**(4), 2089–2102 (2018) 2, 11, 12
11. Kalantari, N.K., Ramamoorthi, R., et al.: Deep high dynamic range imaging of dynamic scenes. ACM Trans. Graph. **36**(4), 144–1 (2017) 1
12. Kim, S., Nam, H., Kim, J., Jeong, J.: C3net: Demoireing network attentive in channel, color and concatenation. In: IEEE Conf. Comput. Vis. Pattern Recog. Worksh. pp. 426–427 (2020) 2, 8, 9, 10
13. Levin, A., Zomet, A., Peleg, S., Weiss, Y.: Seamless image stitching in the gradient domain. In: Eur. Conf. Comput. Vis. pp. 377–389. Springer (2004) 1
14. Li, Y., Brown, M.S.: Exploiting reflection change for automatic reflection removal. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2432–2439 (2013) 2, 10, 11
15. Li, Z., Niklaus, S., Snavely, N., Wang, O.: Neural scene flow fields for space-time view synthesis of dynamic scenes. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 6498–6508 (2021) 2
16. Lin, C.C., Pankanti, S.U., Natesan Ramamurthy, K., Aravkin, A.Y.: Adaptive as-natural-as-possible image stitching. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1155–1163 (2015) 1
17. Liu, B., Shu, X., Wu, X.: Demoir\'eing of camera-captured screen images using deep convolutional neural network. arXiv preprint arXiv:1804.03809 (2018) 8
18. Liu, J., Yang, W., Yang, S., Guo, Z.: Erase or fill? deep joint recurrent rain removal and reconstruction in videos. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3233–3242 (2018) 11, 12, 13

19. Liu, Y.L., Lai, W.S., Yang, M.H., Chuang, Y.Y., Huang, J.B.: Learning to see through obstructions. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 14215–14224 (2020) 2, 7, 11
20. Liu, Y.L., Lai, W.S., Yang, M.H., Chuang, Y.Y., Huang, J.B.: Learning to see through obstructions with layered decomposition. arXiv e-prints pp. arXiv–2008 (2020) 2, 9, 10, 11
21. Lu, E., Cole, F., Dekel, T., Zisserman, A., Freeman, W.T., Rubinstein, M.: Omnimatte: Associating objects and their effects in video. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4507–4515 (2021) 3
22. Martin-Brualla, R., Radwan, N., Sajjadi, M.S., Barron, J.T., Dosovitskiy, A., Duckworth, D.: Nerf in the wild: Neural radiance fields for unconstrained photo collections. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 7210–7219 (2021) 2
23. Meneghetti, G., Danelljan, M., Felsberg, M., Nordberg, K.: Image alignment for panorama stitching in sparsely structured environments. In: Scandinavian Conference on Image Analysis. pp. 428–439. Springer (2015) 2
24. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4460–4470 (2019) 2
25. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: Eur. Conf. Comput. Vis. pp. 405–421. Springer (2020) 2
26. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 165–174 (2019) 2
27. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. In: Int. Conf. Comput. Vis. pp. 5865–5874 (2021) 2
28. Park, K., Sinha, U., Hedman, P., Barron, J.T., Bouaziz, S., Goldman, D.B., Martin-Brualla, R., Seitz, S.M.: Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. ACM Trans. Graph. **40**(6) (dec 2021) 2
29. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 724–732 (2016) 7
30. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10318–10327 (2021) 2
31. Shaham, T.R., Gharbi, M., Zhang, R., Shechtman, E., Michaeli, T.: Spatially-adaptive pixelwise networks for fast image translation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 14882–14891 (2021) 2
32. Sitzmann, V., Martel, J., Bergman, A., Lindell, D., Wetzstein, G.: Implicit neural representations with periodic activation functions. Adv. Neural Inform. Process. Syst. **33** (2020) 1, 3, 8
33. Sitzmann, V., Zollhoefer, M., Wetzstein, G.: Scene representation networks: Continuous 3d-structure-aware neural scene representations. Adv. Neural Inform. Process. Syst. **32**, 1121–1132 (2019) 2
34. Skorokhodov, I., Ignatyev, S., Elhoseiny, M.: Adversarial generation of continuous images. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10753–10764 (2021) 2
35. Tancik, M., Mildenhall, B., Wang, T., Schmidt, D., Srinivasan, P.P., Barron, J.T., Ng, R.: Learned initializations for optimizing coordinate-based neural representations. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2846–2855 (2021) 14

36. Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J., Ng, R.: Fourier features let networks learn high frequency functions in low dimensional domains. In: Adv. Neural Inform. Process. Syst. vol. 33 (2020) 1, 3
37. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Deep image prior. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 9446–9454 (2018) 3
38. Wei, W., Yi, L., Xie, Q., Zhao, Q., Meng, D., Xu, Z.: Should we encode rain streaks in video as deterministic or stochastic? In: Int. Conf. Comput. Vis. pp. 2516–2525 (2017) 2, 11, 12
39. Wronski, B., Garcia-Dorado, I., Ernst, M., Kelly, D., Krainin, M., Liang, C.K., Levoy, M., Milanfar, P.: Handheld multi-frame super-resolution. ACM Trans. Graph. **38**(4), 1–18 (2019) 1
40. Xu, D., Chu, Y., Sun, Q.: Moiré pattern removal via attentive fractal network. In: IEEE Conf. Comput. Vis. Pattern Recog. Worksh. pp. 472–473 (2020) 2, 8, 9, 10
41. Xue, T., Rubinstein, M., Liu, C., Freeman, W.T.: A computational approach for obstruction-free photography. ACM Trans. Graph. **34**(4), 1–11 (2015) 9
42. Yan, Q., Gong, D., Shi, Q., Hengel, A.v.d., Shen, C., Reid, I., Zhang, Y.: Attention-guided network for ghost-free high dynamic range imaging. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1751–1760 (2019) 1
43. Yan, Q., Zhang, L., Liu, Y., Zhu, Y., Sun, J., Shi, Q., Zhang, Y.: Deep HDR imaging via a non-local network. IEEE Trans. Image Process. **29**, 4308–4322 (2020) 1
44. Yang, W., Liu, J., Feng, J.: Frame-consistent recurrent video deraining with dual-level flow. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1661–1670 (2019) 2, 12
45. Zhang, X., Ng, R., Chen, Q.: Single image reflection separation with perceptual losses. In: IEEE Conf. Comput. Vis. Pattern Recog. (2018) 8