Modeling Deep Learning Based Privacy Attacks on Physical Mail

Bingyao Huang, Ruyi Lian, Dimitris Samaras, Haibin Ling
Stony Brook University, NY, USA
{bihuang, rulian, samaras, hling}@cs.stonybrook.edu

Abstract

Mail privacy protection aims to prevent unauthorized access to hidden content within an envelope since normal paper envelopes are not as safe as we think. In this paper, for the first time, we show that with a well designed deep learning model, the hidden content may be largely recovered without opening the envelope. We start by modeling deep learning-based privacy attacks on physical mail content as learning the mapping from the camera-captured envelope front face image to the hidden content, then we explicitly model the mapping as a combination of perspective transformation, image dehazing and denoising using a deep convolutional neural network, named Neural-STE (See-Through-Envelope). We show experimentally that hidden content details, such as texture and image structure, can be clearly recovered. Finally, our formulation and model allow us to design envelopes that can counter deep learning-based privacy attacks on physical mail.

Introduction

With the recent advances in optical devices and deep learning algorithms, traditional paper envelopes may not be as safe as we think. For example, (Redo-Sanchez et al. 2016) show that a *closed* book can be read through using terahertz time-domain spectroscopy; and imaging through scattering media methods (Xin et al. 2019; Satat et al. 2017; Popoff et al. 2010a,b; Kim et al. 2015; Drémeau et al. 2015; Feng et al. 1988; Freund, Rosenbluh, and Feng 1988; Katz, Small, and Silberberg 2012; Katz et al. 2014; Bertolotti et al. 2012; Judkewitz et al. 2015; Yoon et al. 2020) can recover the hidden content behind a volume of refractive media, such as ground glass and human tissue. These methods may be extended and applied to attack sealed physical mail.

In this paper, we show that with a carefully designed deep learning model and sampled images captured in a controlled lab environment (see Fig. 1), the hidden content within the envelope such as the texture and image structure can be recovered without opening it. We start by formulating privacy attacks on physical mail as learning the mapping from the camera-captured envelope image to the hidden content, then we decompose this mapping into a combination of perspective transformation, image dehazing (He, Sun, and Tang 2010; Ren et al. 2016) and deblurring (Ren et al. 2019; Pan et al. 2020). Afterwards, we design three learnable CNN modules to model the three subprocesses. By studying why and how sealed paper mail privacy is subject to deep learning-based attacks, we show that our model can be used to test whether a paper envelope is safe against such attacks and to design safer envelopes in the future.

For the deep learning-based attacks part, to account for perspective transformations in camera image formation, we use WarpingNet, a module similar to a spatial transformer network (STN) (Jaderberg et al. 2015), to warp the camera-captured envelope front face image to the canonical camera frontal view (aligned with the ground truth hidden content). Then, we extend the traditional dehazing formulation (He, Sun, and Tang 2010; Ren et al. 2016) to account for the paper envelope's blur operation, transmittance and surface reflectance under the environment light, such that the camera-captured image is a linear combination of the radiance of blurred hidden content, the transmittance of the paper envelope and the reflected light of the surface. Finally, we incorporate such formulation into a deep neural network and explicitly infer these essential intermediate components using respective CNN modules (*i.e.*, WarpingNet, DehazingNet and RefineNet). This model is trained using sample image pairs of the camera-captured envelope front face (with a hidden printed paper in it) and the ground truth of the printed pattern. In order to counter hypothetical deep learning-based attacks, we first use our model to test the privacy protection ability of an envelope, and then leverage the envelope properties learned by our attack model to design envelopes that are safe against deep learning-based attacks.

Our contributions can be summarized as follows:

- The proposed Neural-STE is the first to model deep learning-based privacy attacks on physical mail.
- Neural-STE is non-trivially designed as a combination of perspective transformation, image dehazing and deblurring.
- Neural-STE can be used to test the privacy-preserving properties of an envelope, and to design envelopes that are safe against deep learning-based attacks.
- We propose the first benchmark for modeling privacy attacks on physical mail. The source code, benchmark dataset and experimental results are publicly available at https://github.com/BingyaoHuang/Neural-STE.

Related Work

Our work is most related to (Redo-Sanchez et al. 2016) that reads through a closed book using terahertz time-domain spectroscopy. The difference is that we aim to model deep learning-based privacy attacks on physical mail without specialized devices such as a terahertz time-domain system. Moreover, we leverage such an attack model to test envelope safety, and to design safer envelopes. The next related work is (Satat et al. 2017) that classifies hidden contents behind the scattering media from the single photon avalanche photodiode (SPAD) captured speckle patterns. It is different from our work since we aim to recover the hidden image rather than classify it. Our work is the first to focus on modeling privacy attacks on physical mail, and to use such a model for mail privacy protection. Another class of related work is imaging through scattering media that aims to recover hidden contents behind a scattering volume. Finally, our learning-based attack method also relates to image dehazing and deblurring as our deep learning-based attack model incorporates the two formulations in network design.

Imaging through scattering media In this paper, we only review studies on occluding scattering media (Yoon et al. 2020), since they relate most to our work. Semi-transparent media such as weather, water, etc., are beyond this paper's scope. Imaging through scattering media can be grouped into traditional optics-based and learning-based methods. Traditional methods use time-resolved measurements (Xin et al. 2019; Satat et al. 2017), transmission matrices (Popoff et al. 2010a,b; Kim et al. 2015; Drémeau et al. 2015) or optical memory effects (Feng et al. 1988; Freund, Rosenbluh, and Feng 1988; Katz, Small, and Silberberg 2012; Katz et al. 2014; Bertolotti et al. 2012; Judkewitz et al. 2015) to reconstruct the hidden scene behind the scattering media. A comprehensive review can be found in (Yoon et al. 2020).

Rather than explicitly modeling the light scattering process, learning-based methods (Horisaki, Takagi, and Tanida 2016; Lyu et al. 2019; Li et al. 2018; Sun, Xia, and Kamilov 2018; Guo et al. 2020; Satat et al. 2017; Li, Xue, and Tian 2018) address this issue as an image-to-image translation problem, *i.e.*, translating the sensor-captured speckle patterns to the appearance of the real hidden contents (or projected virtual objects). The first work of this kind (Horisaki, Takagi, and Tanida 2016) uses pixel-wise support vector regression (SVR) to recover the projected faces behind the scattering media. However, the pixel-wise SVR overfits on faces such that when the testing objects are non-faces the predictions still show strong face patterns. Instead of assuming pixel-wise mapping, deep CNN-based methods (Lyu et al. 2019; Li et al. 2018; Sun, Xia, and Kamilov 2018; Guo et al. 2020; Satat et al. 2017; Li, Xue, and Tian 2018) show better accuracy and generalization.

It is worth noting that our problem is different from imaging through scattering media methods, because sealed pieces of physical mail are different from regular scattering media such as ground glass and human tissue. In addition: 1) imaging through scattering media methods aim to recover accurately the optical properties of a refractive media volume and the hidden contents behind it, while in our setting we focus on recovering empirical properties of paper envelopes and the hidden contents inside them. Moreover, we show how these empirical properties can be used for mail privacy protection; 2) they usually require specialized optical devices such as such as lasers, projectors, beam splitters, polarizers and SPAD, while we only use a DSLR camera and controllable room lights; and 3) as for data collection they use an additional projector to *project* various sampling patterns onto the back of the scattering media to create sample images, and the model input/output image are geometrically registered, while we manually replace the hidden content within the paper envelopes. In our experiment, we show that directly applying a deep learning-based imaging through scattering media method to attack a piece of physical mail may not work well.

Image dehazing Image dehazing (He, Sun, and Tang 2010; Ren et al. 2016; Pan et al. 2020) aims to remove haze and reveal the hidden scene radiance. Due to haze *weak* scattering properties, blur can be ignored and the camera-captured image I can be formulated as a linear combination of the haze-free scene radiance J, the transmission of the haze t and the atmospheric light A:

$$I(x) = J(x)t(x) + A(1 - t(x)) \tag{1}$$

Obviously the equation above cannot be applied to the physical mail attack problem because: (1) Eq. 1 assumes that each pixel x is independent from other pixel radiance values, which only holds when blur is weak. However, unlike haze, paper envelopes can strongly blur the scene radiance, such that each pixel's radiance is a linear combination (blur/convolution) of its neighboring radiance values (see Fig. 1). (2) In Eq. 1, the transmittance t is assumed uniform across RGB channels and the atmospheric light A is assumed uniform and spatially invariant, but for colored and textured materials with spatially variant microstructures (Papas, de Mesa, and Jensen 2014), apparently t and A are spatially nonuniform across RGB channels. Such complexity makes explicitly solving for t and A impossible; instead we can infer their empirical versions from sample images using deep learning-based methods.

Image deblurring In general, image blurring can be formulated as:

$$I = J \otimes h_x + n \tag{2}$$

where I is a camera-captured blurred image, J is the latent clean image, h_x is an unknown spatially variant blur kernel, $\otimes$ is the convolution operator and n is the additive noise. This problem can be solved by imposing certain priors on the blurring kernel and noise, such as assuming a known h_x (Richardson 1972; Lucy 1974; Gonzales and Woods 2002), the sparsity of image gradients (Chan and Wong 1998; Fergus et al. 2006; Levin et al. 2009) and the dark channel prior (Pan et al. 2016). Recently, deep image priors (Ulyanov, Vedaldi, and Lempitsky 2018; Gandelsman, Shocher, and Irani 2019) and GAN-based models (Pan et al. 2020; Kupyn et al. 2018) show clear advantages over previous works. In this paper, inspired by the state-of-the-art learning-based approaches (Ulyanov, Vedaldi, and Lempitsky 2018; Gandelsman, Shocher, and Irani 2019; Pan et al. 2020; Kupyn et al. 2018), we show that it is possible to implicitly solve deblurring in our attack model.

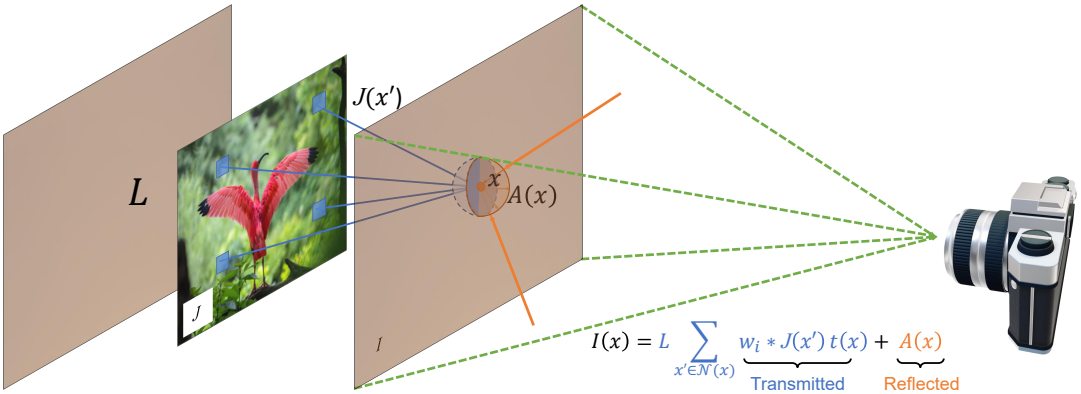

Figure 1: System setup and empirical image formation model. A color printed paper is put within the envelope (distance exaggerated for illustration). Deep learning-based privacy attacks aim to recover the hidden printed paper J from the camera-captured envelope front surface image I. Our formulation models I as a linear combination of the incident environment light L, blurred transmitted paper radiance J and the envelope's front face reflected radiance A. We simplify inter-reflections and subsurface scattering and absorb them in A.

Method

Problem Formulation

Our setup is shown in Fig. 1, where a hidden printed paper is placed within an envelope. Following the notation of (He, Sun, and Tang 2010), denote the camera-captured image as I, the radiance of the hidden printed paper that we aim to recover as J, and the transmittance of the envelope's front face as t. Let A be the reflected (not transmitted) radiance of the envelope's front face under normal environment light (*i.e.*, with room lights on).

Extending Eq. 1 and Eq. 2 to our problem, the radiance of the camera-captured image I at pixel x is given by:

$$\begin{aligned} I(x) &= L \sum_{x' \in \mathcal{N}(x)} \omega_i J(x') t(x) + A(x) \\ &= LJ \otimes h_x t(x) + A(x) \end{aligned} \tag{3}$$

where L is the intensity of the incident environment light on the back face of the hidden content and $\mathcal{N}(x)$ is a neighborhood of x and $\{J(x') | x' \in \mathcal{N}(x)\}$ is a patch of J centered at x. The transmitted radiance at x is the weighted sum of all radiance values above the blue hemisphere of x, *i.e.*, $L \sum_{x' \in \mathcal{N}(x)} \omega_i J(x')$ (Fig. 1). This operation can be approximated by convolving J with a spatially variant convolution kernel h_x, *i.e.*, $LJ \otimes h_x$, where $\otimes$ is the convolution operator. Note that the blurring kernel h_x is also related to the distance between the hidden content and the envelope, *e.g.*, the size of the blurring kernel h_x increases as the distance increases. We assume that L is a constant scalar and it can be absorbed in $t(x)$. Then, the camera-captured radiance is the sum of the transmitted radiance $J \otimes h_x t(x)$ and the reflected radiance $A(x)$ of the envelope.

Clearly the problem in Eq. 3 is highly ill-posed, since the unknown A, t and h_x are hard to obtain. One intuition is to directly estimate J from sample image pairs like (I, J) using an image-to-image translation model (Guo et al. 2020; Isola et al. 2017; Wang et al. 2018), however, such general models are not designed for this problem and tend to obtain suboptimal solutions (see Fig. 3 and Table 1). In this paper, for privacy attacks on physical mail, we decompose this problem as dehazing, deblurring and denoising. First, we estimate the unknowns A, t using a CNN with some constraints, then we explicitly compute the blurred radiance by $J \otimes h_x = (I - A)/t$ as inspired by (He, Sun, and Tang 2010). Finally, we recover the hidden printed paper radiance J using an image refinement network to deblur and denoise and to improve color and texture details.

One unaddressed challenge of mail privacy attack is that the above operations assume that the camera-captured image is aligned with the ground truth image as shown in Fig. 1. However, in practice the hidden content J is real printed paper, manually placed within the envelope, thus there is no guarantee that it is aligned with the camera canonical frontal view. Although some CNN-based image-to-image translation models, *e.g.*, Pix2pix (Isola et al. 2017) can reconstruct both geometry and colors without explicitly modeling the geometry, the output may be subject to a suboptimal solution (color and texture detail loss) when the training samples are limited, as shown in Fig. 3 columns 2-6 (Fig. 3).

To address this issue, we apply a differentiable image warping module to automatically align the input and output images. As shown in our setup Fig. 1, since for mail, both the envelope and the hidden printed paper are approximately planar, a homography is sufficient to correct the geometric distortions. Thus, we explicitly model the perspective transformation by an 8-DoF homography. In summary, our Neural-STE consists of three modules that infer the following intermediate results, respectively:

$$\begin{aligned} A &= \mathcal{F}_A(\mathcal{G}(\mathcal{T}(I))) \\ t &= \mathcal{F}_t(\mathcal{G}(\mathcal{T}(I))) \\ \hat{J} &= \phi((\mathcal{T}(I) - A)/t) \end{aligned} \tag{4}$$

where $\mathcal{T}$ is an STN-like (Jaderberg et al. 2015) homography warping network (Fig. 2) that warps the input image

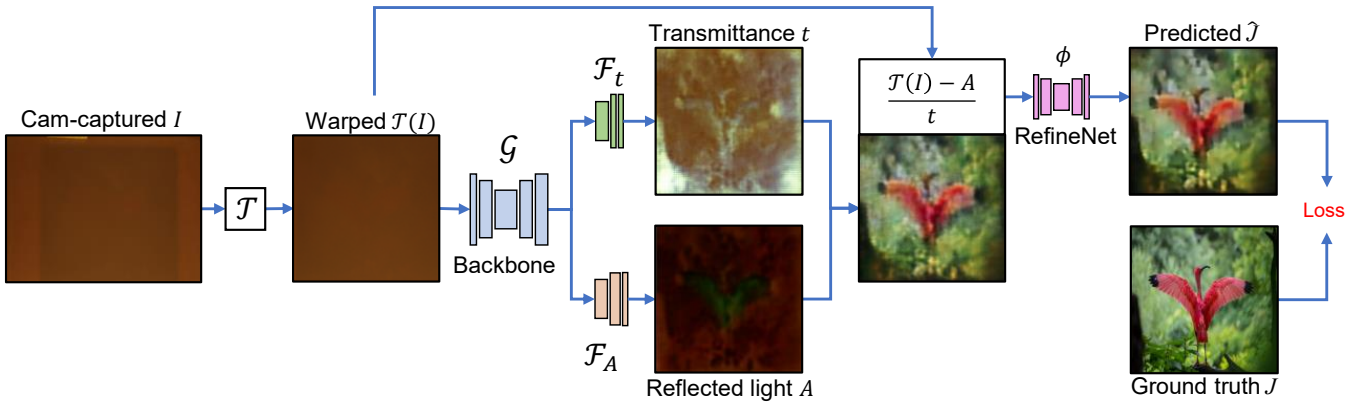

Figure 2: Network architecture of our Neural-STE. It consists of three modules: WarpingNet $\mathcal{T}$, DehazingNet ($\mathcal{G}, \mathcal{F}_A, \mathcal{F}_t$) and RefineNet ϕ. These modules together with the losses allow us to utilize our image formation model to effectively model privacy attacks on physical mail problem.

to the camera canonical frontal view; $\mathcal{G}$ is a feature extraction function, and $\mathcal{F}_A$ and $\mathcal{F}_t$ are two modules that infer the envelope's transmittance t and its surface reflected light A, respectively. ϕ is an image refinement function which includes image deblurring, denoising and color and texture refinement, for conciseness we use image "refinement" in the rest of the paper.

Network Architecture

Given the formulation in Eq. 4, we design our Neural-STE to have three modules (Fig. 2): an STN-based (Jaderberg et al. 2015) **WarpingNet** that warps the camera-captured image to the camera canonical frontal view (*i.e.*, aligned with the ground truth hidden image J); an encoder-decoder backbone network **DehazingNet** to infer transmittance t and surface reflected light A, and compute the coarse dehazed image $(\mathcal{T}(I)-A)/t$; and a **RefineNet** to improve texture and color details of the coarse dehazed image.

WarpingNet ($\mathcal{T}$) is inspired by STN (Jaderberg et al. 2015) and consists of two convolutional layers, two max pooling layers and two fully connected layers. The module firstly infers a 3×3 homography H from the input image, then warps the input image using this predicted homography by $\mathcal{T}(I) = \texttt{imwarp}(I, H)$[1], such that the warped image $\mathcal{T}(I)$ is roughly aligned with the ground truth hidden image J (Fig. 2). Our experimental comparisons show that for our problem, directly learning this geometric transformation is hard if not explicitly modeled as a homography (see **Ours w/o warp** in Fig. 3 and Table 1).

DehazingNet ($\mathcal{G}$, $\mathcal{F}_t, \mathcal{F}_A$) consists of an encoder-decoder backbone network $\mathcal{G}$ and two light weight subnets $\mathcal{F}_t$ and $\mathcal{F}_A$, where the backbone network extracts a 64-dimensional feature map from the input image. Then, we design $\mathcal{F}_t$ as a deconvolutional layer and a convolutional layer followed by a sigmoid activation layer, which predicts the envelope's transmittance t. Afterwards, we design $\mathcal{F}_A$ as two convolutional layers followed by a deconvolutional layer and a sigmoid activation layer. Similarly, it predicts the envelope reflected light A. Finally the coarse hidden printed paper radiance is computed by $J_{\text{coarse}} = (\mathcal{T}(I) - A)/t$.

RefineNet (ϕ). According to our formulation in Eq. 3, recovering the hidden content within a physical mail is more than just dehazing; we need to account for other deformations such as blurring and noise. Note that instead of explicitly estimating the blur kernel, inspired by the success of recently proposed learning-based approaches (Ulyanov, Vedaldi, and Lempitsky 2018; Gandelsman, Shocher, and Irani 2019; Pan et al. 2020; Kupyn et al. 2018), we design **RefineNet** as a CNN with a skip connection (He et al. 2016), such that the color and texture details can be learned as a residual image. In Fig. 4, comparing the refined image $\hat{J} = \phi(J_{\text{coarse}})$ with the coarse image J_{coarse}, we see clearly improved details.

Additional constraints. With the three parameterized modules, training image pairs and a proper loss function $\mathcal{L}$, the network parameters can be learned by:

$$\{\mathcal{T}, \mathcal{G}, \mathcal{F}_t, \mathcal{F}_A, \phi\}_\theta = \arg\min \mathcal{L}(\hat{J} = \phi((\mathcal{T}(I) - A)/t),\ J) \tag{5}$$

However, directly training this network may lead to a sub-optimal solution and the output images may contain strange color or lack color (the 8th-9th columns of Fig. 3). In our setup, we have observed that the surface reflected light A dominates the warped camera-capture image $\mathcal{T}(I)$ (*i.e.*, contributes more light than the transmitted light) due to the ambient light, thus we assume that A should look like $\mathcal{T}(I)$. Then we impose this constraint as a pixel-wise L_2 loss $\|\mathcal{T}(I) - A\|_2^2$. Moreover, the computed coarse result $J_{\text{coarse}} = (\mathcal{T}(I) - A)/t$ should look like the ground truth J, except for some differences in color and texture details, for which we introduce a pixel-wise L_2 loss $\|J - J_{\text{coarse}}\|_2^2$.

Another constraint is that the paper envelopes are not fully opaque ($t \not\approx 0$), otherwise not only the problem is meaningless, but also Eq. 4 may be divided by zero. Thus, we clip the transmittance t to $[0.01, 1]$.

Loss function. As shown in Eq. 6, our loss function consists

[1]Taking MATLAB's `imwarp` as an example.

Model	PSNR↑	RMSE↓	SSIM↑	Model (degraded)	PSNR↑	RMSE↓	SSIM↑
Cam-captured	8.2767	0.6682	0.2695	Ours "black box"	14.1106	0.3442	0.3914
PSDNet (Guo et al. 2020)	12.8952	0.3981	0.3717	Ours w/o refine	11.4025	0.4696	0.3034
Pix2pix (Isola et al. 2017)	12.2620	0.4238	0.3409	Ours w/o warp	14.3415	0.3345	0.4125
Pix2pixHD (Wang et al. 2018)	12.0964	0.4303	0.3193	Ours w/o A con.	14.7582	0.3239	0.4421
Neural-STE (ours)	**15.0275**	**0.3127**	0.4449	Ours w/o J con.	14.9082	0.3151	**0.4460**

Table 1: Quantitative comparison. Results are averaged over three setups, each containing 50 testing images. "Cam-captured" is the similarity between the camera-captured envelope front face and the ground truth. See supplementary for separate measurements for each setup.

of three terms, an image reconstruction loss $\mathcal{L}_{\text{recon}}$ (Eq. 7), *i.e.*, the pixel-wise L_1 + SSIM loss (Zhao et al. 2017) between the predicted hidden printed paper content image $\hat{J}$ and the ground truth J; and the two constraint-based losses above. Then our deep learning-based attack model is trained using Eq. 5 with the loss below.

$$\mathcal{L} = \mathcal{L}_{\text{recon}}(J, \hat{J}) + \|J - J_{\text{coarse}}\|_2^2 + 0.1\,\|A - \mathcal{T}(I)\|_2^2 \quad (6)$$

$$\mathcal{L}_{\text{recon}}(J, \hat{J}) = |J - \hat{J}| + 1 - \text{SSIM}(J, \hat{J}) \quad (7)$$

System configuration and implementation. The proposed setup consists of a Canon 6D camera with the resolution set to 320×240. We color print 500 colorful textured images at US letter size as ground truth hidden contents. Unlike imaging through scattering media, for each capture, we manually replace the printed paper within the envelopes and this operation requires touching the envelopes, thus the shape and pose of the hidden printed paper and the envelopes are inevitably changed each time, making the hidden content recovery more difficult. The distance between the camera and the envelope is around one meter. The only light sources are various room lights. The collected data is available as our Neural-STE dataset. We implement Neural-STE using PyTorch (Paszke et al. 2017) and Kornia (Riba et al. 2019), and optimize it using the Adam optimizer (Kingma and Ba 2015). The initial learning rate and penalty factor are set to 10^{-3} and $5 * 10^{-4}$, respectively. Then, we train the model for 4,000 iterations on three Nvidia GeForce 1080Ti GPUs with a batch size of 16, taking about 18 minutes to train.

Experimental Evaluations

Privacy Attacks on Physical Mail

In this section, we quantitatively and qualitatively evaluate and compare the proposed Neural-STE with PSDNet (Guo et al. 2020), a learning-based image through scattering media method, Pix2pix (Isola et al. 2017), a general GAN-based image-to-image translation model, Pix2pixHD (Wang et al. 2018), an improved version of Pix2pix, and degraded versions of the proposed method.

Evaluation benchmark. We prepared two different sets of envelopes and three different setups, which were configured to cover three levels of difficulty. As shown in Fig. 3, the red box shows a thin kraft envelope imaged under bright room light. The green box shows a thick kraft envelope imaged under bright room light; and the blue box shows a thick kraft envelope under normal room light. For each setup, we split the captured 500 image pairs into 450 training samples and 50 testing samples. Then, the hidden contents $\hat{J}$ recovered by different methods are compared with the ground truth J using PSNR, RMSE and SSIM (Wang et al. 2004).

Since our method is the first to model privacy attacks on physical mail, there is no previous work to compare with. Instead, we compare with PSDNet (Guo et al. 2020), an image through scattering media method. The mapping from the camera-captured image to the hidden content radiance is directly learned without an explicit image formation model like us. While the original PSDNet is designed to only work for grayscale images, we extend it to RGB by increasing the input and output channels. As shown in Fig. 3, PSDNet is unable to recover the hidden contents on the hardest setup (the blue box). Please see the supplementary material for comparisons on grayscale images.

We then compare with a GAN-based general image-to-image translation network Pix2pix (Isola et al. 2017) and its improved version Pix2pixHD (Wang et al. 2018). We train them for 23,000 iterations with a batch size of one. Note that Pix2pix and Pix2pixHD have more parameters than our model, yet they cannot generate satisfactory results as shown in Fig. 3 and Table 1, because they are designed for general image-to-image translation with a relatively large training dataset (both number and diversity), and they may not work well for the privacy attacks on physical mail setting when the training data is limited.

Ablation study. To show the effectiveness of our formulation and network architecture, we compare the proposed Neural-STE with its degraded versions, each with a certain module or constraint removed. For example, **Ours "black box"** is a naive UNet-like (Ronneberger, Fischer, and Brox 2015) model without WarpingNet $\mathcal{T}$, DehazingNet[2] $\mathcal{F}_t, \mathcal{F}_A$ or RefineNet ϕ and the camera-captured image is resized to 256×256 then input to the backbone network $\mathcal{G}$. Moreover, rather than explicitly computing the dehazed and refined image using Eq. 4, the output image is predicted by a three-channel convolutional layer concatenated to the backbone network. **Ours w/o warp** is Neural-STE without WarpingNet $\mathcal{T}$, and the camera-captured image is also resized to 256×256 before fed to the backbone network $\mathcal{G}$. **Ours w/o refine** is the same as Neural-STE but with RefineNet ϕ re-

[2]Note that $\mathcal{F}_t$ and $\mathcal{F}_A$ are concurrently used to dehaze the image, thus we disable them together for this degraded version.

Figure 3: Qualitative comparison. We show results from three different setups, with the easiest in red and the hardest in blue. We show two examples for each setup and the results of different methods are shown in the 2nd to 9th columns. The 1st column are the camera-captured envelope front face I. The 2nd to 4th columns are PSDNet (Guo et al. 2020), Pix2pix (Isola et al. 2017) and Pix2pixHD (Wang et al. 2018), respectively. The 5th to 9th columns are degraded versions of the proposed Neural-STE, as described in Ablation study. Please see supplementary for larger versions of the images and more results.

moved. **Ours w/o A con.** and **Ours w/o J con.** are Neural-STE without the constraints on A (*i.e.*, $0.1\,\|A - \mathcal{T}(I)\|_2^2$ in Eq. 6), and J (*i.e.*, $\|J - J_{\text{coarse}}\|_2^2$ in Eq. 6), respectively.

The experimental comparisons in Table 1 and Fig. 3 clearly show that the proposed Neural-STE outperforms degraded versions that only model part of our image formation process. For example, comparing the 2nd to the 6th columns with the 7th to the 9th columns in Fig. 3, it is clear that explicitly modeling the perspective transformation is important for this problem, because when the inferred image is aligned to the ground truth (especially in the early training stages), the model may focus on refining only the color and texture details, which reduces the probability of falling into local minima early on and improves convergence. Note that **Ours w/o refine** has yellowish output because it cannot fully remove the envelope surface color. **Ours w/o A con.** and **Ours w/o J con.** work well for the easiest envelope, but **Ours w/o J con.** fails to recover hidden scene colors from the other two harder envelopes, and **Ours w/o A con.** generates unwanted bluish pattern for the hardest envelope.

We then show the intermediate results of Neural-STE in Fig. 4. The red and the blue boxes are the easiest and the hardest setups, respectively. The columns are, from left to right, the camera-captured image I, the WarpingNet warped image $\mathcal{T}(I)$, the predicted reflected light A, the predicted transmittance t, the coarse dehazed hidden content radiance $J_{\text{coarse}} = (\mathcal{T}(I) - A)/t$, the final refined results (*i.e.*, $\hat{J} = \phi(J_{\text{coarse}})$), and the ground truth hidden content J. These empirical properties can then be used to design safer envelopes that counters deep learning-based attacks below.

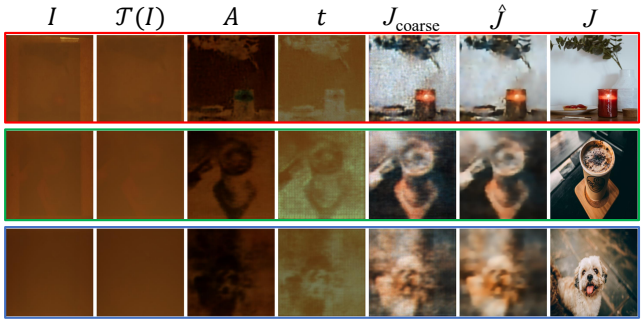

Figure 4: Intermediate results of Neural-STE. See supplementary for more images of the three setups.

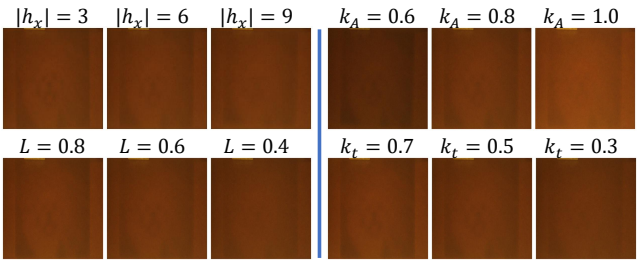

Figure 5: Visualization of the simulated camera-captured image when we tune the controllable parameters, *i.e.*, the size of the blur kernel h_x, the surface reflected light A, the environment light L and the envelope's transmittance t. Here we show an easy setup so that the hidden content is recognizable: as we tune each optical parameter from left to right, the hidden content becomes harder to recognize.

Countering Privacy Attacks on Physical Mail

In this section, we show how to leverage our empirical image formation model and our Neural-STE to design envelopes that can defend mail privacy against deep learning-based attacks. We conducted both simulated and real experiments.

Simulated experiments. We simulated camera-captured images using our empirical image formation model in Eq. 3 and intermediate results of our attack method. The environment light L, the perspective transformation H, the blur kernel size $|h_x|$, the transmittance coefficient[3] k_t, the surface reflected light coefficient k_A were varied in simulation. In addition, we added Gaussian white noise to the blurred image and Poisson noise to the final camera-captured image.

Then, we simulated nine synthetic setups, *i.e.*, for each of the controllable parameters, we generated three variations. In Fig. 5, we show how each controllable parameter affects appearances of the final simulated camera-captured images. Afterwards, we applied our Neural-STE to the simulated dataset; an example result is shown in Fig. 6. See supplementary for quantitative comparisons. We first applied our Neural-STE to examine envelope security. For example, if the envelope can be successfully attacked (as shown in Fig. 6, **Unsafe envelope**), we redesign the envelope using our empirical image formation model, until our Neural-STE fails. In our experiment, we find that, safer envelopes can be manufactured by using a material that has a more reflective surface (*i.e.*, larger A); by increasing the envelope thickness or using a material that absorbs more light (*i.e.*, smaller transmittance t); and by increasing the distance between the hidden contents and the envelope (*i.e.*, a larger blur kernel h_x). For example, in Fig. 6, **Safe envelope** has the following properties: $|h_x| = 17, k_A = 1.0, k_t = 0.1$.

Real experiments. We prepared a new real envelope and attacked it with the proposed Neural-STE, as shown in Fig. 7, **Unsafe envelope**. Then, according to the findings in simulation results, we placed an additional paper layer in the envelope to reduce the transmittance t and to increase the blur kernel h_x, and we named it **Safe envelope**. Afterwards, we attacked it using Neural-STE, and clearly our method failed to reveal the hidden contents within it.

[3]Note that t and A are maps, here we use coefficients (scalars) to control their strengths, *e.g.*, $k_t * t$ and $k_A * A$.

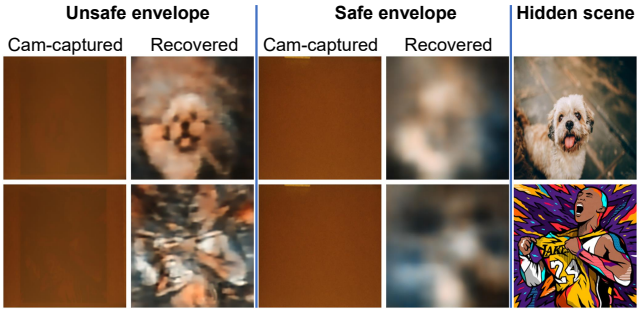

Figure 6: Countering hypothetical deep learning-based privacy attacks on simulated physical mail.

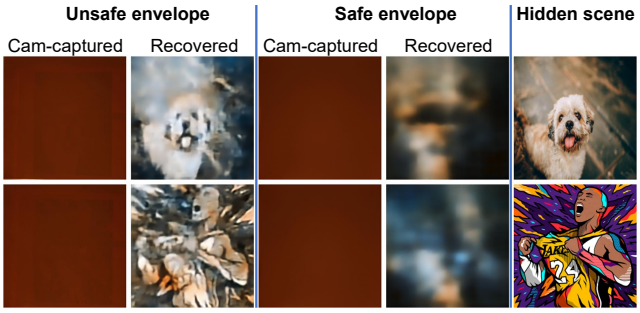

Figure 7: Countering hypothetical deep learning-based privacy attacks on real physical mail.

Conclusion and Limitations

In this paper, we proposed the first deep learning-based method, named Neural-STE for privacy attacks on physical mail. Our method explicitly decomposes an empirical image formation model into a combination of perspective transformation, blurring, transmission and surface reflected light, and then we non-trivially designed respective CNN modules to learn these intermediate results. We proposed the first privacy attacks on physical mail benchmark and we expect it to facilitate future work in this direction. Our experimental results on this benchmark clearly show that normal envelopes are not as safe as we think. Finally, we leverage the empirical image formation model and Neural-STE to design envelopes that can counter such attacks.

Limitations. We have only tested our Neural-STE on kraft envelopes and it may not work well on other materials with much stronger scattering properties. Moreover, we have not tested the method for privacy attacks on folded text, which is much more challenging. Extending this method to (counter) such attacks is definitely an interesting direction to explore.

Acknowledgements. This work was partially supported by the Partner University Fund, the SUNY2020 ITSC, a gift from Adobe, the Yahoo Faculty Research and Engagement Program Award, and the US NSF Grants 2006665.

References

Bertolotti, J.; Van Putten, E. G.; Blum, C.; Lagendijk, A.; Vos, W. L.; and Mosk, A. P. 2012. Non-invasive imaging through opaque scattering layers. *Nature* 491(7423): 232–234.

Chan, T. F.; and Wong, C.-K. 1998. Total variation blind deconvolution. *TIP* 7(3): 370–375.

Drémeau, A.; Liutkus, A.; Martina, D.; Katz, O.; Schülke, C.; Krzakala, F.; Gigan, S.; and Daudet, L. 2015. Reference-less measurement of the transmission matrix of a highly scattering material using a DMD and phase retrieval techniques. *Optics express* 23(9): 11898–11911.

Feng, S.; Kane, C.; Lee, P. A.; and Stone, A. D. 1988. Correlations and fluctuations of coherent wave transmission through disordered media. *Physical review letters* 61(7): 834.

Fergus, R.; Singh, B.; Hertzmann, A.; Roweis, S. T.; and Freeman, W. T. 2006. Removing camera shake from a single photograph. In *SIGGRAPH*, 787–794.

Freund, I.; Rosenbluh, M.; and Feng, S. 1988. Memory effects in propagation of optical waves through disordered media. *Physical review letters* 61(20): 2328.

Gandelsman, Y.; Shocher, A.; and Irani, M. 2019. Double-DIP": Unsupervised image decomposition via coupled deep-image-priors. In *CVPR*, volume 6, 2.

Gonzales, R. C.; and Woods, R. E. 2002. *Digital image processing*. Prentice hall New Jersey.

Guo, E.; Zhu, S.; Sun, Y.; Bai, L.; Zuo, C.; and Han, J. 2020. Learning-based method to reconstruct complex targets through scattering medium beyond the memory effect. *Optics express* 28(2): 2433–2446.

He, K.; Sun, J.; and Tang, X. 2010. Single image haze removal using dark channel prior. *T-PAMI* 33(12): 2341–2353.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Horisaki, R.; Takagi, R.; and Tanida, J. 2016. Learning-based imaging through scattering media. *Optics express* 24(13): 13738–13743.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-Image Translation with Conditional Adversarial Networks. *CVPR* .

Jaderberg, M.; Simonyan, K.; Zisserman, A.; and Kavukcuoglu, K. 2015. Spatial transformer networks. In *NeurIPS*.

Judkewitz, B.; Horstmeyer, R.; Vellekoop, I. M.; Papadopoulos, I. N.; and Yang, C. 2015. Translation correlations in anisotropically scattering media. *Nature physics* 11(8): 684–689.

Katz, O.; Heidmann, P.; Fink, M.; and Gigan, S. 2014. Non-invasive single-shot imaging through scattering layers and around corners via speckle correlations. *Nature photonics* 8(10): 784.

Katz, O.; Small, E.; and Silberberg, Y. 2012. Looking around corners and through thin turbid layers in real time with scattered incoherent light. *Nature photonics* 6(8): 549–553.

Kim, M.; Choi, W.; Choi, Y.; Yoon, C.; and Choi, W. 2015. Transmission matrix of a scattering medium and its applications in biophotonics. *Optics express* 23(10): 12648–12668.

Kingma, D. P.; and Ba, J. 2015. ADAM: A method for stochastic optimization. In *ICLR*.

Kupyn, O.; Budzan, V.; Mykhailych, M.; Mishkin, D.; and Matas, J. 2018. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *CVPR*, 8183–8192.

Levin, A.; Weiss, Y.; Durand, F.; and Freeman, W. T. 2009. Understanding and evaluating blind deconvolution algorithms. In *CVPR*, 1964–1971. IEEE.

Li, S.; Deng, M.; Lee, J.; Sinha, A.; and Barbastathis, G. 2018. Imaging through glass diffusers using densely connected convolutional networks. *Optica* 5(7): 803–813.

Li, Y.; Xue, Y.; and Tian, L. 2018. Deep speckle correlation: a deep learning approach toward scalable imaging through scattering media. *Optica* 5(10): 1181–1190.

Lucy, L. B. 1974. An iterative technique for the rectification of observed distributions. *The astronomical journal* 79: 745.

Lyu, M.; Wang, H.; Li, G.; Zheng, S.; and Situ, G. 2019. Learning-based lensless imaging through optically thick scattering media. *Advanced Photonics* 1(3): 036002.

Pan, J.; Dong, J.; Liu, Y.; Zhang, J.; Ren, J.; Tang, J.; Tai, Y. W.; and Yang, M.-H. 2020. Physics-Based Generative Adversarial Models for Image Restoration and Beyond. *T-PAMI* .

Pan, J.; Sun, D.; Pfister, H.; and Yang, M.-H. 2016. Blind image deblurring using dark channel prior. In *CVPR*, 1628–1636.

Papas, M.; de Mesa, K.; and Jensen, H. W. 2014. A Physically-Based BSDF for Modeling the Appearance of Paper. In *Computer Graphics Forum*, volume 33, 133–142. Wiley Online Library.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in PyTorch. In *NeurIPS-W*.

Popoff, S.; Lerosey, G.; Carminati, R.; Fink, M.; Boccara, A.; and Gigan, S. 2010a. Measuring the transmission matrix in optics: an approach to the study and control of light propagation in disordered media. *Physical review letters* 104(10): 100601.

Popoff, S.; Lerosey, G.; Fink, M.; Boccara, A. C.; and Gigan, S. 2010b. Image transmission through an opaque material. *Nature communications* 1(1): 1–5.

Redo-Sanchez, A.; Heshmat, B.; Aghasi, A.; Naqvi, S.; Zhang, M.; Romberg, J.; and Raskar, R. 2016. Terahertz time-gated spectral imaging for content extraction through layered structures. *Nature communications* 7(1): 1–7.

Ren, D.; Zhang, K.; Wang, Q.; Hu, Q.; and Zuo, W. 2019. Neural blind deconvolution using deep priors. *arXiv preprint arXiv:1908.02197* .

Ren, W.; Liu, S.; Zhang, H.; Pan, J.; Cao, X.; and Yang, M.-H. 2016. Single image dehazing via multi-scale convolutional neural networks. In *ECCV*, 154–169. Springer.

Riba, E.; Mishkin, D.; Ponsa, D.; Rublee, E.; and Bradski, G. 2019. Kornia: an Open Source Differentiable Computer Vision Library for PyTorch. In *WACV*.

Richardson, W. H. 1972. Bayesian-based iterative method of image restoration. *JoSA* 62(1): 55–59.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*. Springer.

Satat, G.; Tancik, M.; Gupta, O.; Heshmat, B.; and Raskar, R. 2017. Object classification through scattering media with deep learning on time resolved measurement. *Optics express* 25(15): 17466–17479.

Sun, Y.; Xia, Z.; and Kamilov, U. S. 2018. Efficient and accurate inversion of multiple scattering with deep learning. *Optics express* 26(11): 14678–14688.

Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. 2018. Deep image prior. In *CVPR*, 9446–9454.

Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In *CVPR*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE TIP* .

Xin, S.; Nousias, S.; Kutulakos, K. N.; Sankaranarayanan, A. C.; Narasimhan, S. G.; and Gkioulekas, I. 2019. A theory of fermat paths for non-line-of-sight shape reconstruction. In *CVPR*, 6800–6809.

Yoon, S.; Kim, M.; Jang, M.; Choi, Y.; Choi, W.; Kang, S.; and Choi, W. 2020. Deep optical imaging within complex scattering media. *Nature Reviews Physics* 1–18.

Zhao, H.; Gallo, O.; Frosio, I.; and Kautz, J. 2017. Loss functions for image restoration with neural networks. *IEEE TCI* .