To Reviewer 1 For "twice more data" case (called DnCNN-SURE*), we treated two different realizations of the same image as different images. Once patches are extracted, we randomly permuted all patches for every epoch and optimized the network with them. Your requests are interesting and we wish to show you the results for your requests, but due to our hardware limitation, training was not working with two times of our current batch size. However, in light of stochastic optimization, we argue that our random permutation does not seem problematic. We wanted to keep our current settings since we would like to use the same optimization setting for all methods (so, needed fixed batch size).

For the imperfect ground truth experiments, SURE requires known σ_{noisy}, but our eSURE requires known σ_{gt} (otherwise, they are not working). In Tables 2, 3, both σ_{noisy} and σ_{gt} are described in the second and third rows, respectively. In practice, there are several methods to accurately estimate noise levels even for real noise [11], [A, B].

[A] S Pyatykh et al. IEEE Trans Im Proc 22(2):687-99 2012 [B] A Abdelhamed et al. CVPR 1692-1700 2018.

Even though Tables 2, 3 may not be practical, they did show the impact of our proposed method for different levels of σ_{gt} explicitly. However, your comment is correct for practical sense, so we did train one deep neural network with varying $\sigma_{gt} \in [1 - 10]$ and $\sigma_{noisy} \in [10.1 - 55]$ for blind color image denoising and tested on images with a fixed noise level (just like Table 1) as shown in the below table. This new experiment yielded consistent results and our proposed eSURE method still outperforms other methods. Hope that the results of this table are convincing to you.

Methods	CBM3D	DnCNN-SURE	DnCNN-N2N	DnCNN-eSURE	DnCNN-MSE
$\sigma = 25$	30.70	30.92	30.73	**31.15**	31.20
$\sigma = 50$	27.38	27.62	27.70	**27.91**	27.93

To Reviewer 2 It is indeed a good idea to be more explicit in some explanations for easier understanding. We will modify some descriptions to reflect your comments such as $\mathbf{y_1}, \mathbf{y_2}$ in Theorem 3 and $\mathbf{y} - \mathbf{x}, \mathbf{z} - \mathbf{x}$ in Eq. 7 that are two noise vectors with zero mean. We thank you for all suggested literature and we will cite them with proper discussions.

To Reviewer 3 We briefly mentioned some cases with correlated pairs of noisy images in the introduction, but here are more explanations on possible cases for correlated pairs of noisy images. We must say that our work may NOT be a direct solution to the following cases, but we do believe that our work should have an important impact on them.

1) **Imperfect noisy ground truth images.** There exist many ground truth datasets available for denoiser training, but many of them are not completely noise-free [H]. There are also cases where obtaining noise-free ground truth images is challenging such as satellite imaging and/or medical imaging (CT, MRI). For example, high radiation dose is required for clean ground truth of X-ray CT and it often takes tens of hours to obtain one volume of high resolution MR. Thus, dealing with imperfect noisy ground truth seems potentially practical for these applications. One may argue that two noise realizations per image for Noise2Noise could be obtained for the above cases, but note that it requires two times larger storage which may not be favorable for the cases such as satellites with limited resources or medical devices collecting large image volumes. Our work showed example experiments for them.

2) **Applications with spatio-temporal resolution trade-off.** As shown in our work, it is always beneficial to reduce the noise level of input images for denoisers for improved performance. Now let us look at one example. Assume that we have a sequence of four images $\mathbf{y_1} \sim \mathbf{y_4}$ per 100 ms with i.i.d Gaussian noise. If one decided to sacrifice temporal resolution by temporal averaging, then there are several ways to do so. For Noise2Noise, to maintain the independence, one may perform moving average like $(\mathbf{y_1} + \mathbf{y_2})/2$ and $(\mathbf{y_3} + \mathbf{y_4})/2$. However, for our eSURE based method, one may have another option for moving average like $(\mathbf{y_1} + \mathbf{y_2})/2$ and $(\mathbf{y_2} + \mathbf{y_3})/2$. The first averaged images are over 100 ms, but the second averages images are over 75 ms. Thus, we argue that our proposed method provides more flexibility than Noise2Noise for the cases with spatio-temporal resolution trade-off. The concept of temporal averaging for better image quality is actually not just our example. Temporal averaging has been often used for some applications in dynamic imaging [C, D]. Recently, there was a work on high speed camera to violate temporal independence [E].

3) Lastly, we hope that our work with more flexibility will motivate more applications with correlated pairs.

[C] KA Mohan et al. IEEE Trans Comp Imag 1(2):96-111 2015 [D] E Gravier et al. IEEE Trans Im Proc 16(4):932-42 2007 [E] Y Lu et al. Phys Rev Lett 122(19):193904 2019

For AWGN concern, first of all, SURE with AWGN has been extended to other noise models such as exponential family [16] and nonparametric models [17], so it is potentially possible that SURE based unsupervised learning could be extended for other noise models. Secondly, several recent works on real noise denoising exploited a heteroscedastic Gaussian model $\mathbf{y} \sim \mathcal{N}(\mathbf{x}, \alpha + \beta\mathbf{x})$ with image generation procedures [F, G] or local AWGN with pixel-shuffle down-sampling [H]. Since our proposed method can deal with heteroscedastic Gaussian / local AWGN models (SURE is a point-wise estimator), our method could be potentially useful for them. This discussion with [F-H] will be added.

[F] S Guo et al. CVPR 1712-22 2019 [G] T Brooks et al. CVPR 11036-44 2019 [H] Y Zhou et al. ArXiv 2019

Lastly, we will report our ablation study to find the relationship between ϵ and σ that was shown in Line 203.